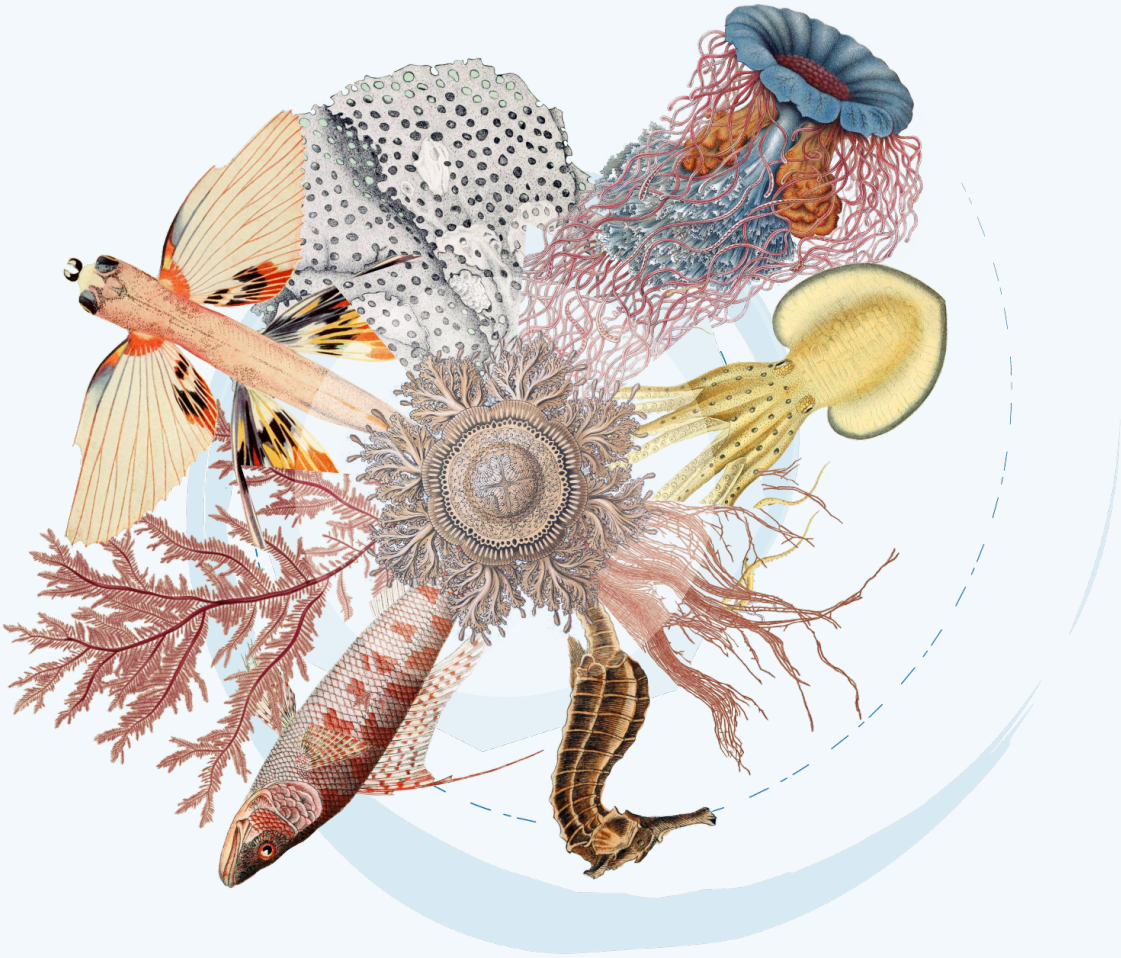

Publishing DNA-derived data through biodiversity data platforms

Kessy Abarenkov • Anders F. Andersson • Andrew Bissett • Anders G. Finstad • Frode Fossøy • Marie Grosjean • Michael Hope • Thomas S. Jeppesen • Urmas Kõljalg • Daniel Lundin • R. Henrik Nilsson • Maria Prager • Pieter Provoost • Dmitry Schigel • Saara Suominen • Cecilie Svenningsen • Tobias Guldberg Frøslev

Version 1.3.0, 7 June 2023

GBIF Global Biodiversity Information Facility

OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM

Table of Contents

Colophon

Suggested citation

Abarenkov K, Andersson AF, Bissett A, Finstad AG, Fossøy F, Grosjean M, Hope M, Jeppesen TS, Kõljalg U, Lundin D, Nilsson RN, Prager M, Provoost P, Schigel D, Suominen S, Svenningsen C & Frøslev TG (2023) Publishing DNA-derived data through biodiversity data platforms, v1.3. Copenhagen: GBIF Secretariat. https://doi.org/10.35035/doc-vf1a-nr22.

Authors

- Kessy Abarenkov, kessy.abarenkov@ut.ee, Natural History Museum and Botanical Garden, University of Tartu, 46 Vanemuise Street, 51003 Tartu, Estonia
- Anders F. Andersson, anders.andersson@scilifelab.se, Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, 17121 Stockholm, Sweden
- Andrew Bissett, Andrew.Bissett@csiro.au, CSIRO O&A, GPO box 1533, Hobart, Tasmania, 7000, Australia
- Anders G. Finstad, anders.finstad@ntnu.no, Department of Natural History, Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
- Frode Fossøy, Frode.Fossoy@nina.no, Centre for Biodiversity Genetics (NINAGEN), Norwegian institute for nature research (NINA), P.O. Box 5685 Torgarden, NO-7485 Trondheim, Norway
- Marie Grosjean, mgrosjean@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Michael Hope, Michael.Hope@ga.gov.au, Atlas of Living Australia, CSIRO National Collections & Marine Infrastructure, GPO Box 1700, Canberra ACT 2601, Australia.
- Thomas S. Jeppesen, tsjeppesen@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Urmas Kõljalg, urmas.koljalg@ut.ee, Natural History Museum and Botanical Garden, University of Tartu, 46 Vanemuise Street, 51003 Tartu, Estonia.
- Daniel Lundin, daniel.lundin@lnu.se, Centre for Ecology and Evolution in Microbial model Systems - EEMiS, Linnaeus University, SE-39182 Kalmar, Sweden
- R. Henrik Nilsson, henrik.nilsson@bioenv.gu.se, University of Gothenburg, Department of Biological and Environmental Sciences, Box 461, 405 30 Göteborg, Sweden
- Maria Prager, maria.prager@scilifelab.se, Science for Life Laboratory, Department of Ecology, Environment and Plant Sciences, Stockholm University; Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet
- Pieter Provoost, p.provoost@unesco.org, Ocean Biodiversity Information System, Jacobsenstraat 1, 8400 Oostende, Belgium
- Dmitry Schigel, dschigel@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Saara Suominen, s.suominen@unesco.org, Ocean Biodiversity Information System, Jacobsenstraat 1, 8400 Oostende, Belgium
- Cecilie Svenningsen, csvenningsen@gbif.org, Global Biodiversity Information Facility, Universitetsparken 15, 2100 København Ø, Denmark
- Tobias Guldberg Frøslev, tfroeslev@gbif.org, Global Biodiversity Information Facility,

Universitetsparken 15, 2100 København Ø, Denmark

Contributors

Valuable discussions with members of ELIXIR, iBOL, GGBN, GLOMICON, and OBIS networks contributed to compilation of this draft. We are especially grateful for inputs and encouragement from Andrew Bentley, Matt Blissett, Pier Luigi Buttigieg, Kyle Copas, Camila A. Plata Corredor, Gabriele Dröge, Torbjørn Ekrem, Birgit Gemeinholzer, Quentin Groom, Tim Hirsch, Donald Hobern, Hamish Holewa, Corinne Martin, Raissa Meyer, Chris Mungall, Daniel Noesgaard, Corinna Paeper, Tim Robertson, Maxime Sweetlove, Andrew Young, John Waller, Ramona Walls, John Wieczorek, Lucie Zinger who have contributed to the GBIF community review process.

License

Persistent URI

https://doi.org/10.35035/doc-vf1a-nr22

Document control

Version 1.3.0 released on 7 June 2023.

This version adds paragraph on marine datasets and the Ocean Biodiversity Information System (OBIS) along with some minor text edits throughout.

Abstract

When genetic information is used to describe or classify a taxon, most users will foresee its use in the context of molecular ecology or phylogenetic research. It is important to realize that a sequence with coordinates and a timestamp is a valuable biodiversity occurrence which is useful in a much broader context than its original purpose. To realize this potential, DNA-derived data needs to be discoverable through biodiversity data platforms. This guide will teach you the principles and approaches of exposing “sequences with dates and coordinates” in the context of broader biodiversity data. The guide covers choices of particular schemas and terms, common pitfalls and good practice, without going into platform-specific details. It will benefit anyone interested in better exposure of DNA-derived data through general biodiversity data platforms, including national biodiversity portals.

Preface

The work on this guide started from discussions at the biodiversity_next conference in 2019 consolidating inputs from various resources, such as:

- Final Report on Environomics Future Science Platform Project
- ALA blog post eDNA records now available on ALA
- Environmental DNA (eDNA) in the ALA
- ALA eDNA data template

- Norwegian Criteria for depositing eDNA samples and data, including vouchered specimens
- Molecular biodiversity data in SBDI, Sweden
- GBIF resources (How) can I publish molecular/sequence/DNA based data to GBIF?
- Molecular data in GBIF
- GBIF quick guide to publishing data and detailed guides to publishing data
- How to publish data to GBIF, as well as DwC/extension field overview.
- Genomic Biodiversity Interest Group

1. Introduction

1.1. Rationale

The last 20 years have brought an increased understanding of the immense power of molecular methods for documenting the diversity of life on earth. Seemingly lifeless and mundane substrates such as soil and sea water turn out to abound with life–although perhaps not in a way that the casual observer may immediately appreciate. DNA-based studies have shown that organism groups such as fungi, insects, oomycetes, bacteria and archaea are everywhere, although we often cannot observe them physically (Debroas et al. 2017). The benefits of molecular methods are not restricted to the microscopic world: there are many organisms, such as some fish species, which can at least theoretically be observed physically but for which it is very costly, labour-intensive, and perhaps invasive to do so (Boussarie et al. 2018). In such situations, DNA data enable us to record the presence (and past presence) of these organisms non-invasively and with minimal effort. These developments mean that we do not always need tangible, physical manifestations of all organisms present at some site in order to record them. All organisms, whether or not they are physically observable, may be important when it comes to understanding biodiversity, ecology and biological conservation.

DNA-derived data enable us to record inconspicuous or otherwise unobservable taxa that fall below the radar of vetted protocols for field work, checklists, depositions into natural science collections, etc. The current maturity of DNA methodologies enables us to record the presence of these organisms to a level of detail that exceeds that of macroscopic observations of organisms in general. However, bearing in mind that DNA methodologies comes with their own problems and biases, it is important to use this moment to define and agree how we should record and report on an organism as present in some substrate or locality through molecular data. Doing so will help avoid significant inefficiencies that have been reported in other domains, in which the lack of standards and guidance has led to very heterogeneous and largely incomparable bodies of data (Berry et al. 2021; Leebens-Mack et al. 2006; Yilmaz et al. 2011; Nilsson et al. 2012; Shea et al. 2023). Moreover, clear documentation of the computational processing from raw sequence reads to deduced species observation, will enable reanalysis when improved methods appear.

DNA-derived occurrence data of species should be as standardized and reproducible as possible, regardless of whether or not the detected species have formal scientific names. In some cases, such occurrence records will hint at previously unknown geographical and ecological properties of described species, thus enriching our body of knowledge on these taxa. In other cases, the data may allow us to amalgamate and visualize information on currently undescribed species, potentially speeding up their eventual formal description. The ability to collect usable data even for unnamed species adds significantly to the many ways in which GBIF and other biodiversity data platforms index the living world, and make this knowledge available to all and for a variety of purposes, including biodiversity conservation. Recent estimates suggest that at least 85 per cent of all extant species are undescribed (Mora et al. 2011; Tedesco et al. 2014). Existing data standards have been designed for

the minority of taxa that have been described. Good practices for dealing with DNA-derived data will help to characterize occurrences of all organisms, whether described or not.

This guide sets out the ways in which DNA-derived occurrence data should be reported for standardized inclusion in GBIF and other biodiversity data platforms. It does not express any view on the issue of access and benefit sharing for digital sequence information, the subject of extensive discussion through the Convention on Biological Diversity (CBD). However, it is worth noting that genetic barcodes and metabarcodes are typically genes or non-coding DNA fragments, which are not suitable for commercial exploitation. As the archiving of sequences through International Nucleotide Sequence Database Collaboration (INDSC) is a widespread norm in sequence-based research, publication of occurrence data originating from sequences does not involve publishing new sequences. In most cases these have already been placed in a public genetic repository. This guide therefore addresses the added value possible from deriving spatio-temporal occurrence data and dna-based names from dna data, rather than the value of the genetic information itself. In addition to dealing with sequence-derived data, this guide also includes suggestions for publishing occurrence data of species derived from qPCR or ddPCR analyses.

Reporting DNA-derived occurrences in an open and reproducible way brings many benefits: notably, it increases citability, highlights the taxa concerned in the context of biological conservation and contributes to taxonomic and ecological knowledge. Additionally, it also provides a mechanism to store occurrence records of undescribed species. When this yet to be described taxon is finally linked to a new Linnaean name, all these linked occurrence records will be immediately available. Each of these benefits provides a strong rationale for professionals to adopt the practices outlined in this guide, helping them to highlight a significant proportion of extant biodiversity, hasten its discovery and integrate it into biological conservation and policy-making.

1.2. Target audiences

This guide has been developed for multiple target audiences: students planning a first DNA-based study, researchers with old sequences and abundance tables they want to revive or preserve, biodiversity data specialists who are new to DNA-derived occurrences, and bioinformaticians familiar with sequence data but new to biodiversity data platforms. The guide does not directly target users of molecular data in biodiversity data platforms, but such users may find section 1.7 on Outputs particularly interesting. The authors' intention is to provide guidance on publishing data and associated attributes from genetic sequence through general biodiversity data platforms.

The flowchart outlines the processing steps involved in publishing amplicon-derived molecular biodiversity data in repositories such as GBIF and national biodiversity data platforms, including those built on the ALA platform. This guide's focus is primarily on the steps following the arrival of raw FASTQ sequences from the sequencing step. By familiarizing themselves with the flowchart—and noting any steps that appear familiar or unclear—users will be able to navigate the content included in the guide.

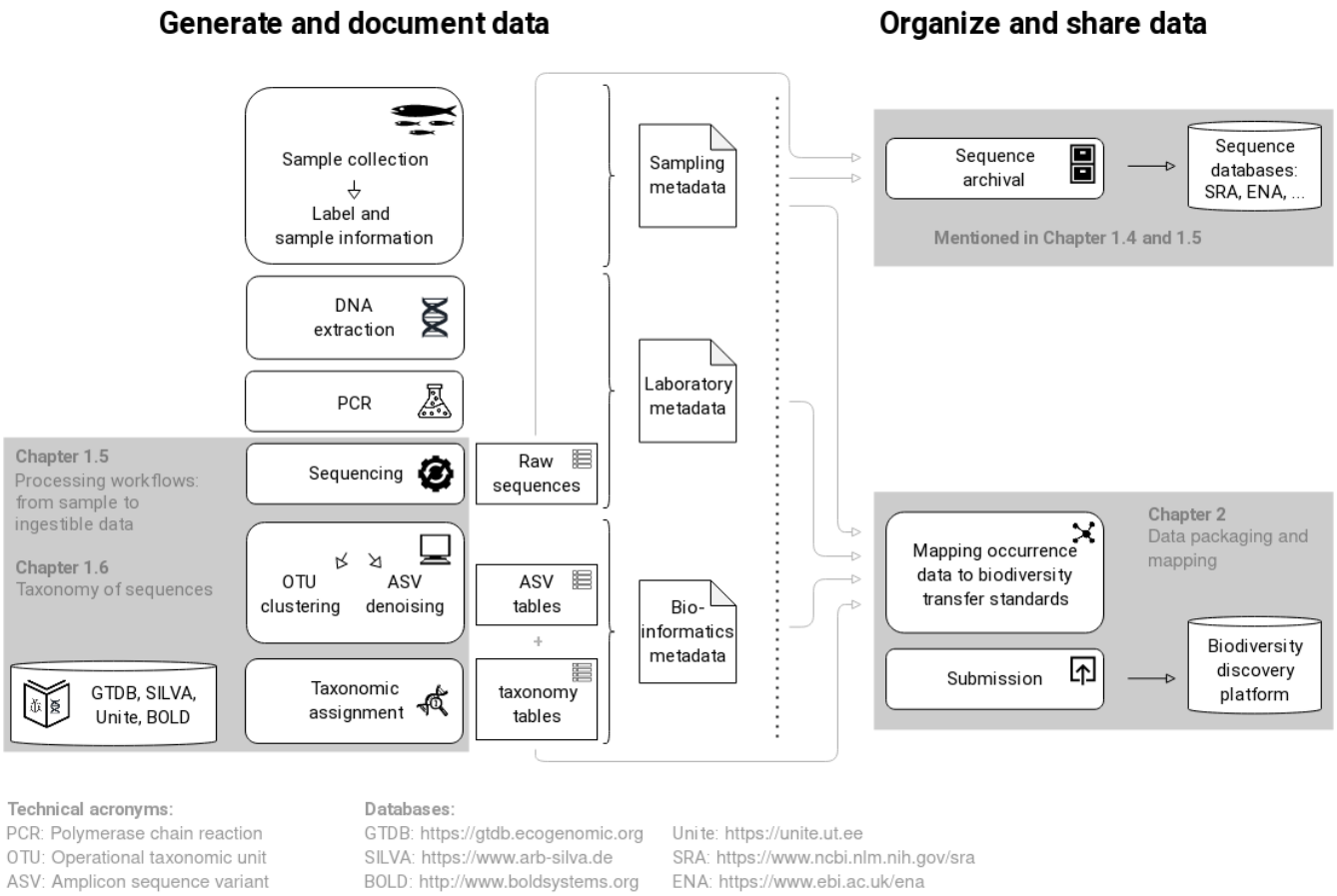

Figure 1. Overall workflow for DNA sequence-derived biodiversity data as described in this guide.

We have done our best to present the information in this guide so that it is useful for each of the audiences described above, but background reading (e.g. GBIF quick guide to data publishing) may be required in certain cases.

1.3. Introduction to DNA-derived occurrence data

DNA-derived biological occurrence data include information derived from DNA from individual organisms, but also from environmental DNA (eDNA, i.e. DNA extracted from environmental samples, Thomsen & Willerslev 2015) and from bulk samples comprising many individuals (e.g. plankton samples or Malaise trap samples consisting of multiple individuals from many species). Currently, the greatest volume of DNA-derived occurrence data derives from eDNA. Since analytical methods and end products are largely similar for all sample sources, the discussion below will focus on eDNA (§ 2.1.1 and § 2.1.2), noting that the outline is applicable to the other sources. Surveys often utilize targeted sequencing of taxonomically and phylogenetically informative genetic markers, but can also use, for example, qPCR-based approaches that do not directly result in DNA sequence data (§ 2.1.3 and § 2.2.2). This guide may appear heavy in DNA related terms; if this is the case, please consult the Glossary.

1.3.1. Environmental DNA as a source for DNA-derived occurrence data

Environmental DNA has been in use as a term since 1987, when it was first used to describe DNA from microbes in sediment samples (Ogram et al. 1987). eDNA is now more broadly used to describe a complex mix of DNA from different organisms (Taberlet et al. 2018 and 2012). Thus, eDNA includes all DNA extracted from a specific environmental sample, regardless of substrate and which species it contains. It may be extracted from a wide range of sources, including skin and hair cells, saliva, soil, faeces, and from living or recently dead organisms (Pietramellara et al. 2009). Environmental DNA often sufficiently represents all organisms in a given sample. In practice, however, the presence of

DNA in the environmental sample depends on an organisms habitat selection, body size, morphology and activity level. Also, the sampling methods used to capture the DNA (Taberlet et al. 2018) and the stage of DNA degradation can affect the presence of DNA.

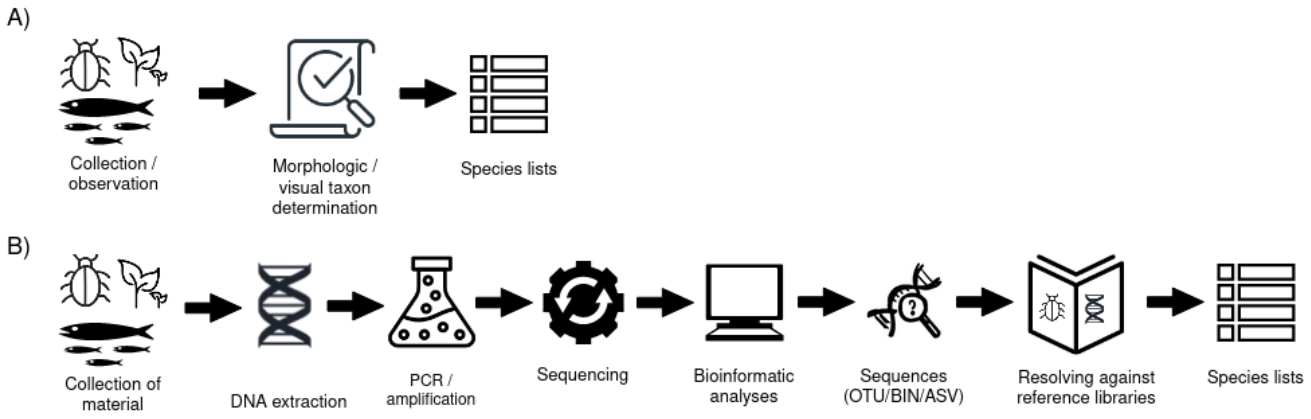

Figure 2. Caricature of sampling processes comparing data collection by A) traditional ecological/biodiversity sampling methods, and B) eDNA-based studies, here exemplified by metabarcoding. This is a simplified representation. For eDNA, most of the steps up to sequencing will involve technical or biological replications to identify contamination and false positives as well as false negative results, making the structure of data and metadata hierarchical. However, studies will often include both types of sampling. For example, if the 'Reference Library' used in B) does not contain all relevant species from a given group of organisms, it will be necessary to go back to A). It may also be that 'Resolving against Reference Library' produced unexpected or unlikely results, in which case further studies using traditional methodology will be required to determine whether the species identified by bioinformatic analysis can be verified.

eDNA is thus a sample type, not a method, including DNA derived from any environmental sample rather than from the capture and sequencing of a targeted individual. Such sample types includes water, soil, sediment and air, but also gut content samples and tissue (plant/animal) where the host DNA is not targeted (Taberlet et al. 2018). A number of analytical methods exist for studying environmental DNA. These can be divided into two main classes: 1) those which aim to detect a specific organism and 2) those which describe an assemblage or a community of a range of organisms. Different methods of analysis will generate different types and volumes of data. Most often DNA concentrations are low, and technical and biological replicates should be used to validate species detection.

Several studies show that, for water samples, analyses based on eDNA may have a higher probability of finding rare and hard to survey species than conventional methods (Thomsen et al. 2012; Biggs et al. 2015; Valentini et al. 2016; Bessey et al. 2020). The same may be true in other environments. Therefore, eDNA may be suitable for monitoring rare red list species and undesirable alien species that often have low densities and that are difficult to detect with conventional methods because sometimes DNA traces can still be detected, although the actual organism is no longer present there. Environmental DNA methods are able to detect cryptic organisms, especially those that are small and unable to be detected by the naked eye (e.g. bacteria and fungi). In addition, eDNA can also be used for observation of many species simultaneously, and may describe entire biological communities or major components of them (Ekrem & Majaneva 2019).

Some studies show a relationship between the amount of DNA for a given species in an environmental sample and the biomass of the species in the environment. One can therefore potentially also think of environmental DNA allowing a so-called semi-quantitative estimate (indirect target) for organism biomass, both from environmental samples and bulk samples (Takahara et al. 2012; Thomsen et al. 2012; Andersen et al. 2012; Ovaskainen et al. 2013; Lacoursière-Roussel et al. 2016; Thomsen et al. 2016; Valentini et al. 2016; Fossøy et al. 2019; Yates et al. 2019; Doi et al. 2017).

However, other studies show little correlation between environmental DNA quantity and estimated population density (Knudsen et al. 2019). PCR, quantification, mixing and other biases are frequently debated. For example, moult, reproduction and mass death can contribute to increased levels of crustacean environmental DNA in water, while turbidity and poor water quality reduce the amount of detectable environmental DNA (Strand et al. 2019). Therefore we encourage data publishers to supply both read counts for each OTU or ASV per sample as well as total read count per sample, as this is necessary information for users to make their own conclusions on presence/absence and (relative) abundance.

1.3.2. DNA-metabarcoding: sequence-derived data

The generation of sequence-derived data is currently increasing fast due to the development of DNA-metabarcoding. This method utilizes general primers to generate thousands to millions of short DNA-sequences for a given group of organisms with the help of high-throughput sequencing (HTS, alt. next-generation sequencing (NGS)). By comparing each DNA-sequence to a reference database such as GenBank (Benson et al. 2006), BOLD (Ratnasingham et al. 2007), or UNITE (Nilsson et al. 2019), each sequence can be assigned to a species or higher rank taxon identity. DNA-metabarcoding is used for samples originating from both terrestrial and aquatic environments, including water, soil, air, sediments, biofilms, plankton, bulk samples and faces, simultaneously identifying hundreds of species (Ruppert et al. 2019).

The identification and classification of organisms from sequence data and marker-based surveys depends on access to a reference library of sequences taken from morphologically identified specimens that are matched against the newly generated sequences. The efficacy of classification depends on the completeness (coverage) and the reliability of reference libraries, as well as the tools used to carry out the classification. These are all moving targets, making it essential to apply taxonomic expertise and caution in the interpreting results (§ 1.6). Availability of of all verified amplicon sequence variants (Callahan et al. 2017) allow for precise reinterpretation of data, intra-specific population genetic analyses (Sigsgaard et al. 2019) and is likely to increase identification accuracy, and for this reason we recommend to share (unclustered) ASV data.

1.3.3. Metagenomic: sequence-derived data

Sequence derived diversity data may also be generated using amplification free metagenomic methods whereby all DNA in a sample is targeted for sequencing (Tyson & Hugenholtz 2005), rather than specific amplicons or barcodes, as described above. Sequence derived diversity data obtained from metagenomic sequencing can be in the form of sequence matches to annotated gene databases (as above) or as (near) complete metagenome assembled genomes (MAGs). While metabarcoding methods still dominate in terms of sequence derived diversity information, metagenomic data is becoming more important, as evidenced by the growing number of MAGS and their utility in informing phylogeny and taxonomy (Parks et al. 2020); discussion of the rapidly evolving methods associated with metagenome analysis is beyond the scope of this document. This document uses metabarcoding as the model for discussion around concepts and methods for publishing sequence derived diversity data, and while the bioinformatic pathways will differ for metagenomic data, the end result (a sequence, often in the form of a contig/assembly) is congruent with the concepts suggested for metabarcoding data (i.e., sample specific, sample collection, data generation and processing workflow metadata should be captured).

1.3.4. qPCR/ddPCR: occurrence data

For the detection of specific species in eDNA-samples, most analyses include species-specific primers and qPCR (Quantitative Polymerase Chain Reaction) or ddPCR (Droplet-Digital Polymerase Chain Reaction). These methods do not generate DNA-sequences, and the occurrence data are completely dependent on the specificity of the primers/assays. Hence, there are strict

recommendations for how to validate such assays and the requirements for publishing data (Bustin et al. 2009, Huggett et al. 2013), as well as the readiness for assays in routine monitoring (Thalinger et al. 2020). Analyses of eDNA-samples using qPCR requires few resources and can be done in most DNA-laboratories. The first example of using eDNA water samples utilized qPCR for detecting the invasive American Bullfrog (*Rana catesbeiana*) (Ficetola et al. 2008), and qPCR analyses of eDNA water samples are regularly used for detecting specific species of fish, amphibians, molluscs, crustaceans and more, as well as their parasites (Hernandez et al. 2020, Wacker et al. 2019, Fossøy et al. 2019, Wittwer et al. 2019). eDNA-detections using qPCR thus generate important occurrence data for single species.

1.4. Introduction to biodiversity publishing

Publishing biodiversity data is largely a process of making species occurrence data findable, accessible, interoperable and reusable, in accordance with the FAIR principles (Wilkinson et al. 2016). Biodiversity data platforms help expose and discover genetic sequence data as biodiversity occurrence records alongside other types of biodiversity data, such as museum collection specimens, citizen science observations, and classical field surveys. The structure, management and storage for each original data source will vary according to the needs of each community. The biodiversity data platforms support data discovery, access and reuse by making these individual datasets compatible with each other, addressing taxonomic, spatial and other inconsistencies in the available biodiversity data. Making data available through single access points supports large-scale data-intensive research, management, and policy. The compatibility between datasets is reached through the process of standardization.

A number of data standards are in use for general biodiversity data (https://www.gbif.org/standards), and a separate set of standards for genetic sequence data (see MIxS and GGBN). This guide reflects some ongoing efforts to increase the compatibility between standards for general biodiversity and genetic data. Standards often highlight the subsets of fields which are most important or most frequently applicable. These subsets may be referenced as “cores”. The preferred format for publishing data in the GBIF and ALA networks is currently the Darwin Core Archive (DwC-A) using the Darwin Core (DwC) data standard. In practice, this is a compressed folder (a zip file) containing data files, in standard comma- or tab-delimited text format, a metadata file (eml.xml) that describes the data resource, and a metafile (meta.xml) that specifies the structure of files and data fields included in the archive. Standardized packaging ensures that the data can travel between systems using specific data exchange protocols. Section 2 of this guide provides recommendations for the mapping of the data files, while guidelines and tools for constructing the xml files can be found here: TDWG, GBIF, and ALA.

A central part of the standardization process is the mapping of fields, which is required to transform the original field (column) structure in a source-data export into a standard field structure. Standardization may also affect the content of the individual fields within each record, for example, by recalculating coordinates to a common system, rearranging date elements, or mapping the contents of fields a standard set of values, often called a vocabulary. The process of standardization also provides an opportunity to improve data quality, for example, by filling in omissions, correcting typos and extra spaces and handling inconsistent use of fields. Such improvements enhance the quality of data and increase its suitability for reuse, but at the same time, data published in any state are better than data that remain unpublished and inaccessible. Standartization is typically applied to a copy or to an export from the data source, leaving the original untouched.

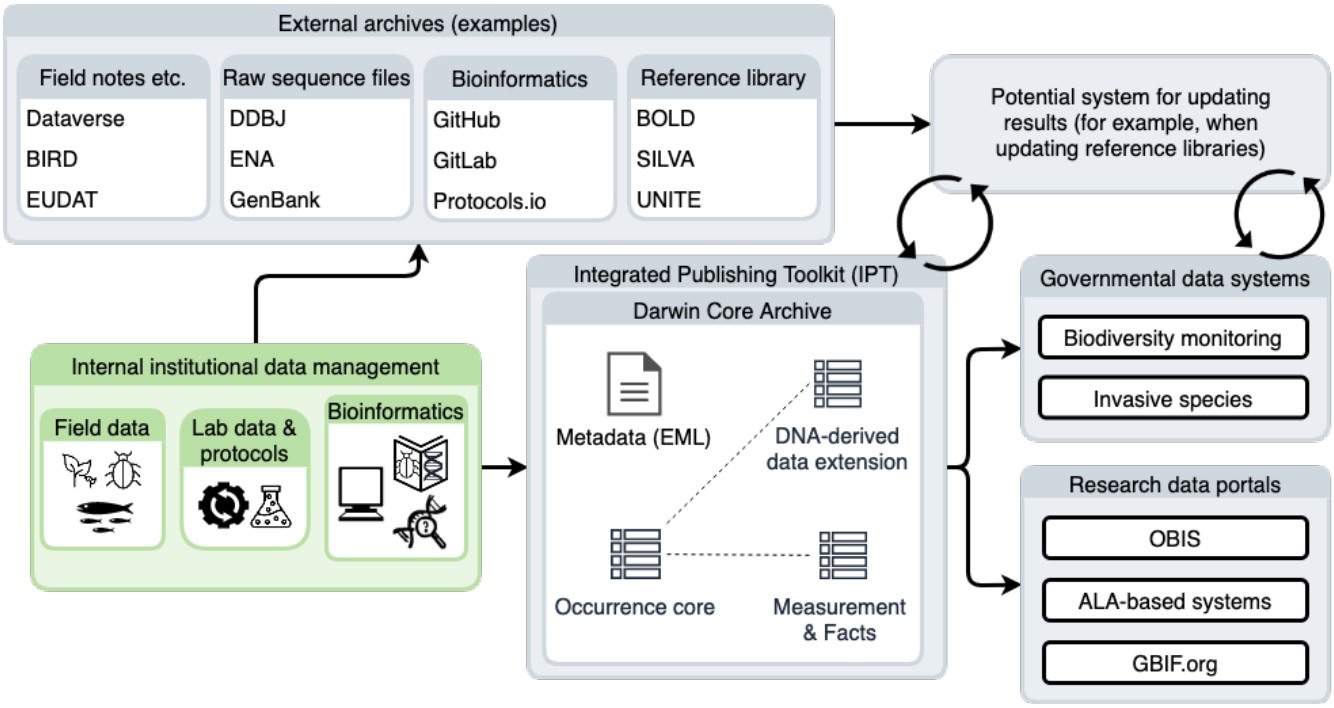

Figure 3. Outline of a platform for reporting and publishing DNA sequences and associated metadata (green box) based on existing systems and data standards (grey boxes). An envisioned system for regular (based on machine-to-machine reading of data) update of results (white box) can either read and update the Darwin Core Archive or various administration systems. The data transfer between the various elements (black arrows) will require various degrees of data transformation and harmonization and may include either mechanical or human quality assessment.

Once a dataset has been through these standardization and data quality processes, it should be placed in an accessible online location and associated with relevant metadata. Metadata–data or information about the dataset includes key parameters that describe the dataset and further improve its discoverability and reuse. Metadata should include other important elements such as authorship, Digital Object Identifiers (DOIs), organizational affiliations and other provenance information, as well as procedural and methodological information about how the dataset was collected and curated. We encourage to provide a description of workflow details and versions including quality control in the methods section in the EML file.

Datasets and their associated metadata are indexed by each data portal: this process enables users to query, filter and process data through APIs and web portals. Unlike journal publications, datasets may be dynamic products that go through multiple versions, with an evolving number of records and mutable metadata fields under the same title and DOI.

Note that holders of genetic sequence data are expected to upload and archive genetic sequence data in raw sequence data repositories such as NCBI's SRA, EMBL's ENA or DDBJ. The sequence archival topic is not covered here, but e.g. Penev et al. (2017) provide a general overview of the importance of data submission and guidelines in connection with scientific publication. Biodiversity data platforms such as ALA, GBIF, and most national biodiversity portals are not archives or repositories for raw sequence reads and associated files. We do, however, stress the importance of maintaining links between such primary data and derived occurrences in Section 2.

1.5. Processing workflows: from sample to ingestible data

Metabarcoding data can be produced from a number of different sequencing platforms (Illumina, PacBio, Oxford Nanopore, Ion Torrent, etc.) that rely on different principles for readout and generation

of data that differ with respect to read length, error profile, whether sequences are single or paired-end, etc. Currently the Illumina short-read platform is the most widely adopted and as such is the basis of the descriptions here. However, the bioinformatics processing of the data follows the same general principles (QC, denoising, classification) regardless of the sequencing technology used (Hugerth et al. 2017, Figure 2).

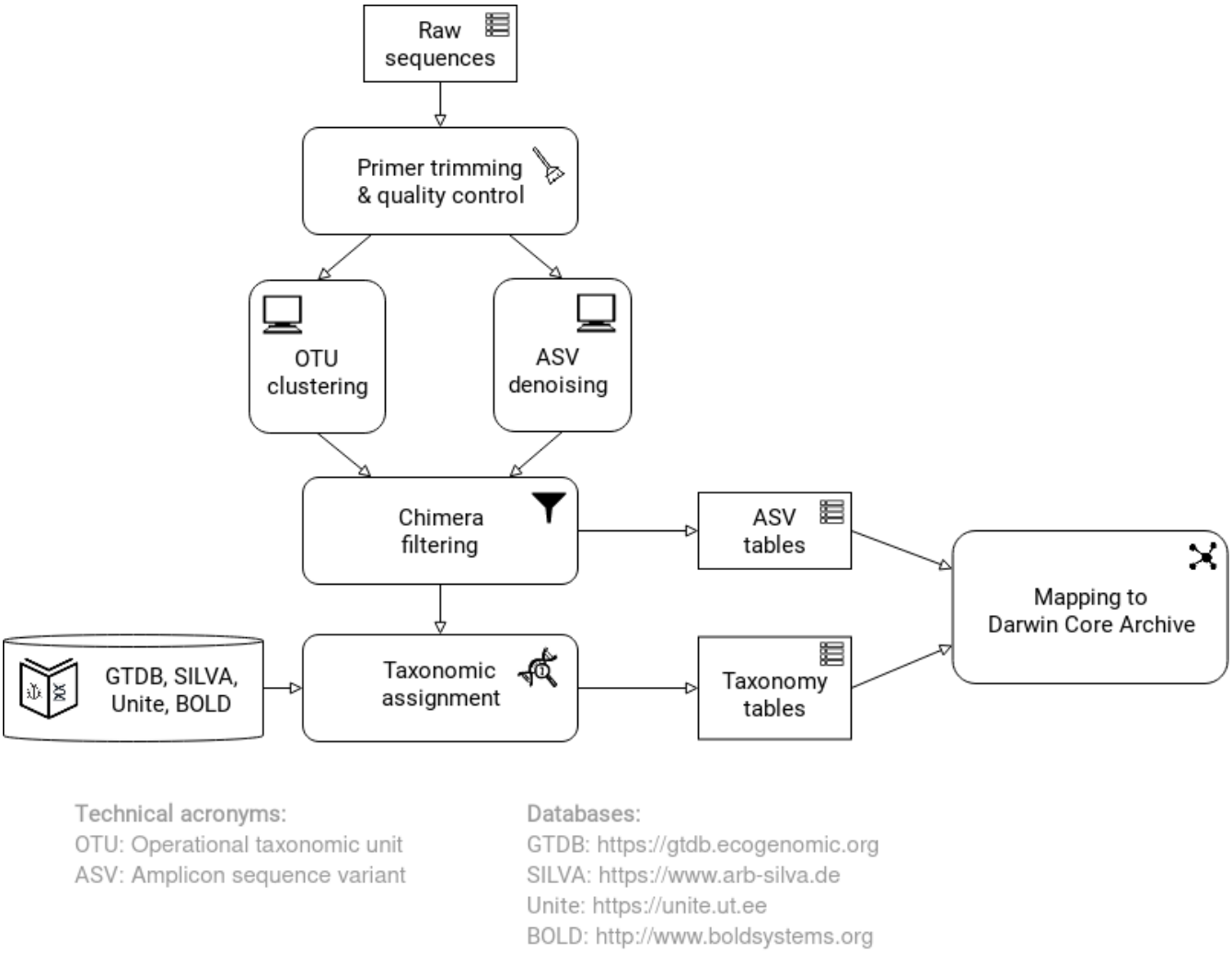

Figure 4. Outline of bioinformatic processing of metabarcoding data.

Typically, the DNA sequences are first pre-processed by removing primer sequences and, depending on the sequencing method used, low quality bases, usually toward the 5' and 3' sequence ends. Sequences not fulfilling requirements on length, overall quality, presence of primers, tags etc. are removed.

The pre-processed sequences can then be assigned a taxon by comparing them against reference databases. When reference databases are incomplete, sequences classification can be done without taxonomic identifications, either by clustering sequences into operational taxonomic units based on their similarity (OTUs; Blaxter et al. 2005) or by denoising the data, i.e. explicitly detecting and excluding PCR/sequencing errors sequences to produce amplicon sequence variants (ASV; also referred to as zero radius OTU (zOTU)). Denoising attempts to correct errors that have been introduced in the PCR and/or sequencing steps, such that the denoised sequences are the set of unique biologically real sequences present in the original sequence mixture. In case of paired-end sequences, the forward and reverse sequences may be denoised separately before merging or else merged prior to denoising. ASVs in the resulting set can differ by as little as one base which is indicative of inter- or intraspecific sequence variation. Operationally, ASVs may be thought of as OTUs without defined radius and while denoising algorithms are typically very good, they do not entirely remove the problems of over-splitting or lumping sequences.

The PCR used for generating the sequencing library can result in the generation of artefactual sequences in the form of chimeras; a single sequence that originates from multiple parent sequences. Such sequences can be detected bioinformatically and removed, and this is typically done after denoising.

Finally, the pre-processed sequences, OTUs or ASVs, are taxonomically classified by comparing them to a database of annotated sequences (often referred to as reference libraries, see § 1.6). As with the previous steps, several alternative methods are available. Most of these are either based on aligning the metabarcoding sequences to the reference sequences or on counting shared k-mers (short exact sequences).

Several open source tools and algorithms exist for bioinformatic processing of metabarcoding data (QIIME2 (Bolyen et al. 2019), DADA2 (Callahan et al. 2016), SWARM (Mahé et al. 2014), USEARCH (Edgar 2010), Mothur (Schloss et al. 2009), LULU (Frøslev et al. 2017), PROTAX (Somervuo et al. 2016), VSEARCH (Rognes et al. 2016)). Given the existence of many popular and well used workflows, we make some recommendations below on analysing data for submission to biodiversity data platforms. This is not to suggest that these are the best methods or most appropriate for all purposes but is an attempt to encourage submission of relatively standardized data that may readily be compared via the platforms. If possible, a well documented and maintained workflow should be used (e.g. nf-core/ampliseq pipeline). Metadata should include workflow details and versions either in the metadata method steps or as a reference in the SOP field in the DNA derived data extension (see mapping in Table 4). Sequence data should be deposited in an appropriate nucleotide archive (NCBI's SRA: Leinonen et al. 2011) or EMBL's ENA (Amid et al. 2020)) and data submitted to the biodiversity platform should include the biosample ID obtained from the archive (see data mapping in § 2.2). Making use of these sample IDs will reduce the chances of duplication and ensure sequence data are readily obtainable should opportunities for re-analysis arise, as reference libraries and bioinformatic tools improve. The core end-product of these pipelines is typically a file of counts of individual OTUs or ASVs in each sample along with the taxonomy assigned to these. This is generated either in tabular format or in the BIOM format (McDonald et.al 2012). OTU or ASV sequences are also often provided in the FASTA format (Pearson & Lipman 1988).

1.6. Taxonomy of sequences

Taxonomic annotation of sequences is a critical step in the processing of molecular biodiversity datasets, as scientific names are key to accessing and communicating information about the observed organisms. The accuracy and precision of such sequence annotation will depend on the availability of reliable reference databases and libraries across all branches of the tree of life, which in turn will require joint efforts from taxonomists and molecular ecologists. Public sequence databases should always be used knowingly of the fact that they suffer from various shortcomings related to, e.g., taxonomic reliability and lack of standardized metadata vocabularies (Hofstetter et al. 2019; Durkin et al. 2020).

Species, as described by taxonomists, are central to biology and attempts at characterizing biodiversity may therefore make use of the end products of taxonomic research. However, unlike DNA sequence data, taxonomic outputs are not always readily amenable to direct algorithmic or computational interpretation: classical taxonomy is a human-driven process which includes manual steps of taxon delimitation, description and naming, culminating in a formal publication in accordance to the international Codes of Nomenclature. As discussed in previous chapters, DNA sequence-based surveys are very effective at detecting hard to observe species and will often identify the presence of organisms currently outside traditional Linnaean taxonomic knowledge. While these guidelines do not address the publication of alternative species checklists derived from sequence data, the disconnection between traditional taxonomy and eDNA efforts is undesirable. Therefore we offer the following recommendations to readers of this guide.

As taxonomy is central to the discovery of biodiversity data, it is highly recommended that any eDNA sequencing efforts should seek to include relevant taxonomic expertise in their study. It will similarly be beneficial if eDNA sequencing studies are able to allocate a portion of their budget to generation and release of reference sequences from previously unsequenced type specimens or other important reference material from the local herbarium, museum, or biological collection. Taxonomists, too, can contribute towards this goal by always including relevant DNA sequences with each new species description (Miralles et al. 2020) and by targeting the many novel biological entities unravelled by eDNA efforts (e.g. Tedersoo et al. 2017).

Most current biodiversity data platforms are organized around traditional name lists and taxonomic indexes. As DNA sequence-derived occurrences are rapidly becoming a significant source of biodiversity data, and as official taxonomy and nomenclature for such data lags, it is recommended that data providers and platforms should continue to explore and include more flexible representations of taxonomy into their taxonomic backbones. These new representations include molecular reference databases (e.g., GTDB, BOLD, UNITE) that recognize sequence data as reference material for previously unclassified organisms. Additionally, we suggest other commonly used molecular databases (e.g., PR2, RDP, SILVA) should develop stable identifiers for taxa and make reference sequences available for those taxa, to allow their use as taxonomic references.

In contrast to classical taxonomy, which is a heavily manual process, clustering DNA sequences into taxonomic concepts relies on algorithmic analysis of similarity and other signals (such as phylogeny and probability), as well as some human editing. The resulting OTUs vary in stability, presence of reference sequences and physical material, alignments and cut-off values, and OTU identifiers such as DOIs (Nilsson et al. 2019). Even more importantly, they vary in scale, from local study- or project-specific libraries to global databases that enable broader cross-study comparison. In contrast to the centralization and codification of Linnaean taxa that are formally described in research publications, OTUs are distributed across multiple evolving digital reference libraries that differ in taxonomic focus, barcode genes and other factors. By associating standard sequences with identified reference specimens, BOLD and UNITE are establishing an essential mapping layer for linking ASVs and OTUs with Linnaean taxonomy. The GBIF backbone taxonomy includes identifiers for UNITE Species Hypotheses (SHs) as well as Barcode Index Numbers (BINs) which allows indexing of species occurrence data taxonomically annotated at the OTU level for primarily Fungi and Animals (GBIF secretariat 2018, Grosjean 2019).

Algorithms for taxonomic annotation of eDNA will typically assign each unique sequence to the nearest taxonomic group in a reference set, based on some criteria for relatedness and confidence. For poorly known groups of organisms, such as prokaryotes, insects and fungi, the annotation may be a non-Linnaean placeholder name for a (cluster-based) taxon (i.e. the ID/number of the relevant SH or BIN), and this taxon may represent a species or even a taxonomic unit above the species level. No reference database contains all species in a given group due to the many unknown, unidentified, and undescribed species on earth. Frequent ignorance of this fact has been the source of numerous taxonomic misidentifications during the last 30 years.

During import into the biodiversity platform (e.g. GBIF or OBIS), the taxonomic resolution for these DNA-based occurrences may be reduced even further, as the names/IDs obtained from comparing with the reference database (e.g. UNITE, BOLD) may not all be included in the taxonomic index of that platform at the time of publication. However, the inclusion of the underlying OTU or ASV sequence for each record will allow future users to potentially identify the sequence to a greater level of granularity, particularly as reference libraries improve over time. Therefore we also recommend to publish all sequences in a study - also those that are presently fully unclassified – as they may well be possible to identify with improved reference databases. In cases where the underlying sequence cannot be included as part of the submitted data, we advocate deposition of a (scientific or placeholder) name of the taxon (e.g. the BOLD BIN or UNITE SH) plus an MD5 checksum of the sequence as a unique taxon ID (see § 2.2, “Data mapping”). MD5 checksums are unidirectional hash

algorithms commonly used for verifying file integrity. In this case, they would provide a unique and repeatable representation of the original sequence that would nevertheless not allow the sequence itself to be recovered. This may be required in cases where sensitivity exists around access. MD5 checksums enable efficient query to determine whether the same exact sequence has been recovered in other eDNA efforts, but it is not a complete replacement of the sequence as MD5s do not enable further analyses. Two sequences differing by even a single base will get two completely different MD5 checksums, such that BLAST-style sequence similarity searches will not work.

1.7. Outputs

The purpose of exposing DNA-derived data through biodiversity platforms is to enable reuse of these data in combination with other biodiversity data types. It is very important to keep this reuse in mind when preparing your data for publication. Ideally, the metadata and data should tell a complete story in such a way that new, uninformed users can use this evidence without any additional consultations or correspondence. Biodiversity data platforms provide search, filtering, browsing, visualizations, data access, and data citation functionality. For metabarcoding data we encourage users to configure filters for minimum absolute and relative read abundance to make a suitable filtering of data. Singletons or any occurrence with an absolute read count below some selected value can be filtered out by setting a minimum read abundance per OTU or ASV (using the field organismQuantity). Occurrences with a relative read abundance below a selected threshold can be filtered out by setting a minimum value of relative organism quantity, which is calculated from the detected reads (organismQuantity) and total reads in the corresponding sample (sampleSizeValue) (§ 2.2.1). Users can often choose data-output formats (e.g. DwC-A, CSV) and then process, clean and transform data into the shape and format needed for the analyses.

At GBIF.org or through the GBIF API, registered users can search, filter, and download biodiversity data in the following three formats:

- **Simple**: a simple, tab-delimited format which includes only the GBIF-interpreted version of the data, as a result of the indexing process. This is suitable for quick tests and direct import into spreadsheets.
- **Darwin Core Archive**: richer format that includes both the interpreted data and the original verbatim version provided by the publisher (prior to indexing and interpretation by GBIF). Because it includes all the metadata and issue flags, this format provides a richer view of the downloaded dataset.
- **Species list**: a simple table format that includes only an interpreted list of unique species names from a dataset or query result.

Regardless of the selected format, each GBIF user download receives a reusable link to the query and a data citation that includes a DOI. This DOI-based citation system provides the means of recognizing and crediting uses to datasets and data originators, improving both the credibility and transparency of the findings based on the data. It is essential to follow data citation recommendations and use DOIs, as good data citation culture is not only the academic norm, but also a powerful mechanism for crediting acknowledging and, therefore, incentivizing data publishers.

2. Data packaging and mapping

This chapter focuses on practical details on turning your data export into a dataset indexed by a biodiversity data platform. § 2.1 will help you understand what is the optimal mapping schema for your data at hand. § 2.2 describes these mappings in detail.

This guide combines the standards for general biodiversity data publishing with genetic DNA-derived

biodiversity data (Figure 5). This “do-section” stops at providing mapping recommendations for different types of DNA-derived data.

Data packaging and publishing pathways vary from platform to platform and are described in general documentation. One of the widespread ways to package data files is currently DwC-A, where data tables are arranged in a star schema, with records (rows) in peripheral extension files pointing to a single record in the central core file (Figure 5). The different types of core files (e.g. occurrence and sampling-event) correspond to different classes of datasets. Although DNA-derived datasets often are event-based in nature, i.e. hundreds or even thousands of quantified sequence occurrences may derive from a single sampling event and thus share most metadata attributes, the current recommendation is to publish data as Occurrence core (Category I or II) with the DNA derived data extension. This approach compensates for limitations of the DwC star schema, which would not allow any occurrence-level data in extension files (such as processed barcode sequences) to point to records in an event core file. We do, however, recommend including an eventID for each core record, to indicate the association between occurrences derived from the same sampling event.

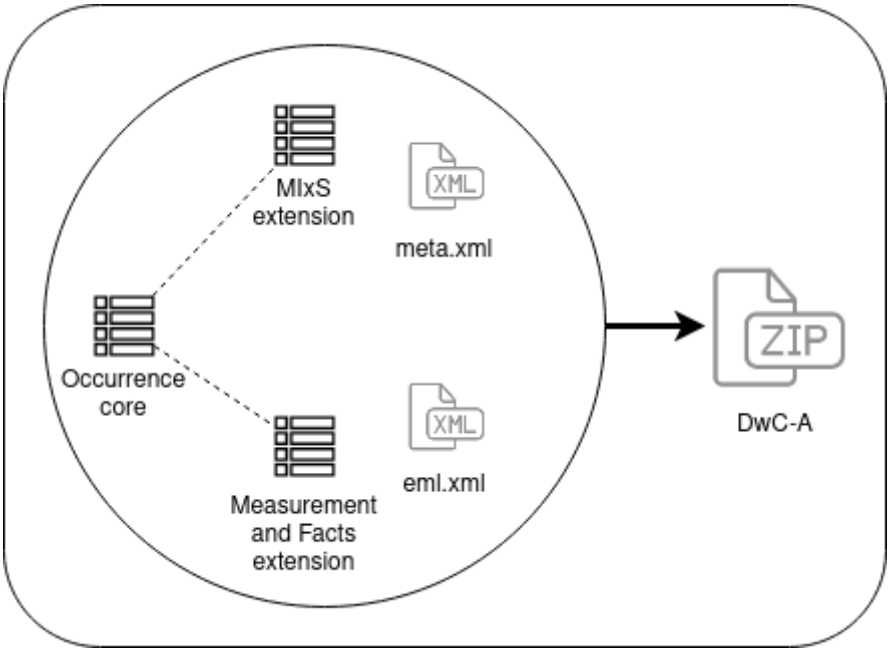

Figure 5. Zoom in of DwC-A / IPT from figure 3 in chapter 1.2. The choice of core entity is mainly a matter of fitting data to the data import mechanism (ingestion) of the biodiversity data platforms. Most data could be formulated as either Occurrence, Event or Taxon core, but as only the core can have extensions, this will affect the choice. It is for example not possible to extend occurrences with DNA sequences if data are packaged using Event core.

2.1. Categorization of your data

For the purpose of this guide, we categorize data into five categories, linked by a key ID field (*eventID*), equivalent to the standards for general biodiversity data, and include fields relevant for DNA-derived data (see § 2.2, “Data mapping”). These five categories seek to reflect the most commonly used molecular approaches to biodiversity characterization and are I) DNA-derived occurrences, II) enriched occurrences, III) targeted species detection, IV) name references and V) metadata only. Examine the decision tree and proceed to the correct section below.

Table 1. A decision tree for DNA-derived data categorization.

Is your data (meta)barcoding or qPCR based?	
(Meta)barcoding ↓	qPCR ↓

Does data consist of digitized genetic material, or sequences, associated with location and time?				**Category III** Targeted species detection
Yes ↓ *Is the genetic material the **only** evidence of a given organism or community?*		No ↓ *Is the dataset a list of DNA-based names?*		
Yes ↓ **Category I** DNA-based occurrences	No ↓ **Category II** Enriched occurrences	Yes ↓ **Category IV** Name references	No ↓ **Category V** Metadata-only	

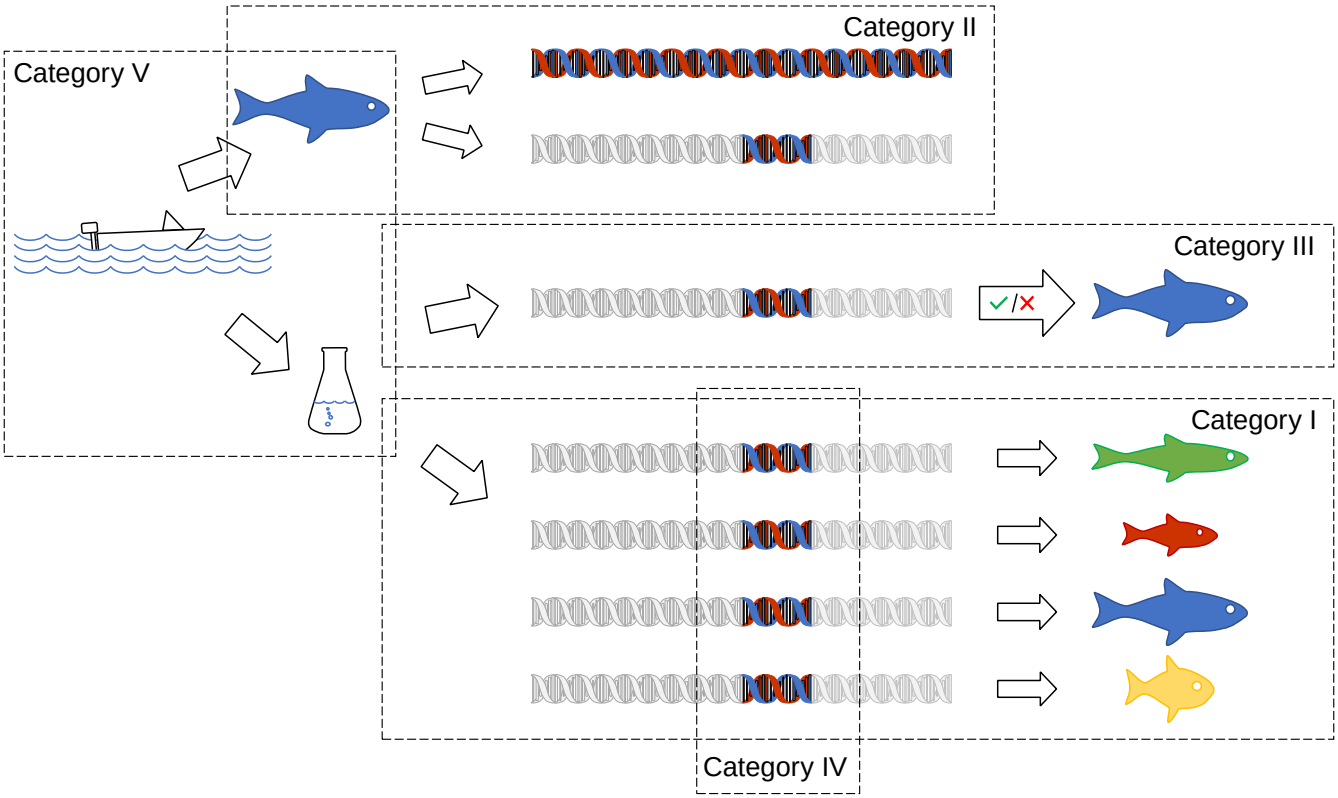

Figure 6. Visual representation of categories I-V.

2.1.1. Category I: DNA-derived occurrences

This category concerns data where a DNA sequence or detection through PCR is the only evidence for the presence of a given organism or community. In other words, the data cannot be traced back to an observable specimen. This is the case for many metagenomics, metabarcoding and eDNA studies.

Examples of DNA-derived occurrence datasets

- MGnify (2019) Impact of rainforest transformation on phylogenetic and functional diversity of soil prokaryotic communities in Sumatra (Indonesia). Sampling event dataset https://doi.org/10.15468/osp7hi accessed via GBIF.org on 2020-04-16.
- MGnify (2020) Marine metagenomes from the bioGEOTRACES project. Sampling event dataset https://doi.org/10.15468/oifcho accessed via GBIF.org on 2020-04-16.
- Bessey C, Jarman SN, Berry O et al. (2020) Maximizing fish detection with eDNA metabarcoding. Environmental DNA: 1–12. https://doi.org/10.1002/edn3.74 (Atlas of Living Australia website at

https://collections.ala.org.au/public/show/dr14581. Accessed 24 June 2020)

For guidance on how to format and share these datasets, see § 2.2.1. General guidelines for Darwin Core occurrence datasets are also available through the DwC-A template for occurrence datasets and Data quality requirements for occurrences.

2.1.2. Category II: Enriched occurrences

If some genetic material is, or can be, associated with an observation or a specimen, we will categorize this type of data as "enriched occurrences". In this context, the sequences are not the only evidence of occurrences. One can always trace the information back to a vouchered specimen or observed organism. This category includes barcoding datasets and some DNA metabarcoding datasets with reference material for example. For more guidance on barcoding, follow Centre for Biodiversity Genomics, University of Guelph (2021).

Examples of Enriched occurrence datasets

- The International Barcode of Life Consortium (2016) International Barcode of Life project (iBOL). Occurrence dataset https://doi.org/10.15468/inygc6 accessed via GBIF.org on 2020-04-16.
- Takamura K (2019) Chironomid Specimen records in the Chironomid DNA Barcode Database. Version 1.9. National Institute of Genetics, ROIS. Occurrence dataset https://doi.org/10.15468/hxhow5 accessed via GBIF.org on 2020-04-16.
- Bessey C, Jarman SN, Stat M, Rohner CA, Bunce M, Koziol A, Power M, Rambahiniarison JM, Ponzo A, Richardson AJ & Berry O (2019) DNA metabarcoding assays reveal a diverse prey assemblage for Mobula rays in the Bohol Sea, Philippines. Ecology and Evolution 9 (5) 2459-2474. https://doi.org/10.1002/ece3.4858, (Atlas of Living Australia website at https://collections.ala.org.au/public/show/dr11663. Accessed 24 June 2020)

For guidance on how to format and share these datasets, see § 2.2.1. General guidelines for Darwin Core occurrence datasets are also available through the DwC-A template for occurrence datasets and Data quality requirements for occurrences.

2.1.3. Category III: Targeted species detection (qPCR/ddPCR)

This category concerns data where a specific (qPCR/ddPCR) assay is used to detect the presence (or absence) of a DNA sequence specific to the target organism in an environmental sample. In this case the occurrence record may not even contain sequence data, as it is the process itself that determines the occurrence. With qPCR/ddPCR analyses for targeted species detection, many studies also report absence of that specific species for a given sample. Absence data is highly dependent on the detection limit of the specific assay, as well as field and lab protocols. As for DNA-metabarcoding data there is an issue of both false negatives and false positives, and it is important that sufficient information is reported for evaluating the records.

Examples of targeted species occurrence datasets

- Strzelecki, Joanna; Feng, Ming; Berry, Olly; Zhong, Liejun; Keesing, John; Fairclough, David; Pearce, Alan; Slawinski, Dirk; Mortimer, Nick. Location and transport of early life stages of Western Australian Dhufish Glaucosoma hebraicum. Floreat, WA: Fisheries Research and Development Corporation; 2013. http://hdl.handle.net/102.100.100/97533 (Atlas of Living Australia website at https://collections.ala.org.au/public/show/dr8131. Accessed 22 July 2020)

For guidance on how to format and share these datasets, see § 2.2.2. General guidelines for Darwin Core occurrence datasets are also available through the DwC-A template for occurrence datasets and Data quality requirements for occurrences.

2.1.4. Category IV: Name references

This category corresponds to DNA-derived names, derived from clustering or denoising (error-correction based models), such as stable non-Linnaean Operational Taxonomic Units (OTU), Amplicon Sequence Variant (ASV) and Barcode Index Numbers (BIN)—in other words, any reference to taxa or provisional names that are defined outside of the Linnaean taxonomy. Numerous projects produce local project- or study-specific libraries of OTUs, and although it is technically possible to publish these as checklists, they have limited to no value for data linking or interpretation; as a result, we do not encourage their publication through biodiversity data platforms. However, the inclusion of the widely adopted, stable, global, digitally referenceable OTUs into Linnaean taxonomic backbones is critically important for indexing unnamed “dark” biodiversity. GBIF have accumulated experience in integrating such large and global reference libraries of OTUs into the GBIF taxonomic backbone, which allows the display of OTUs under the nearest parent taxon which has a Scientific name (Figure 7).

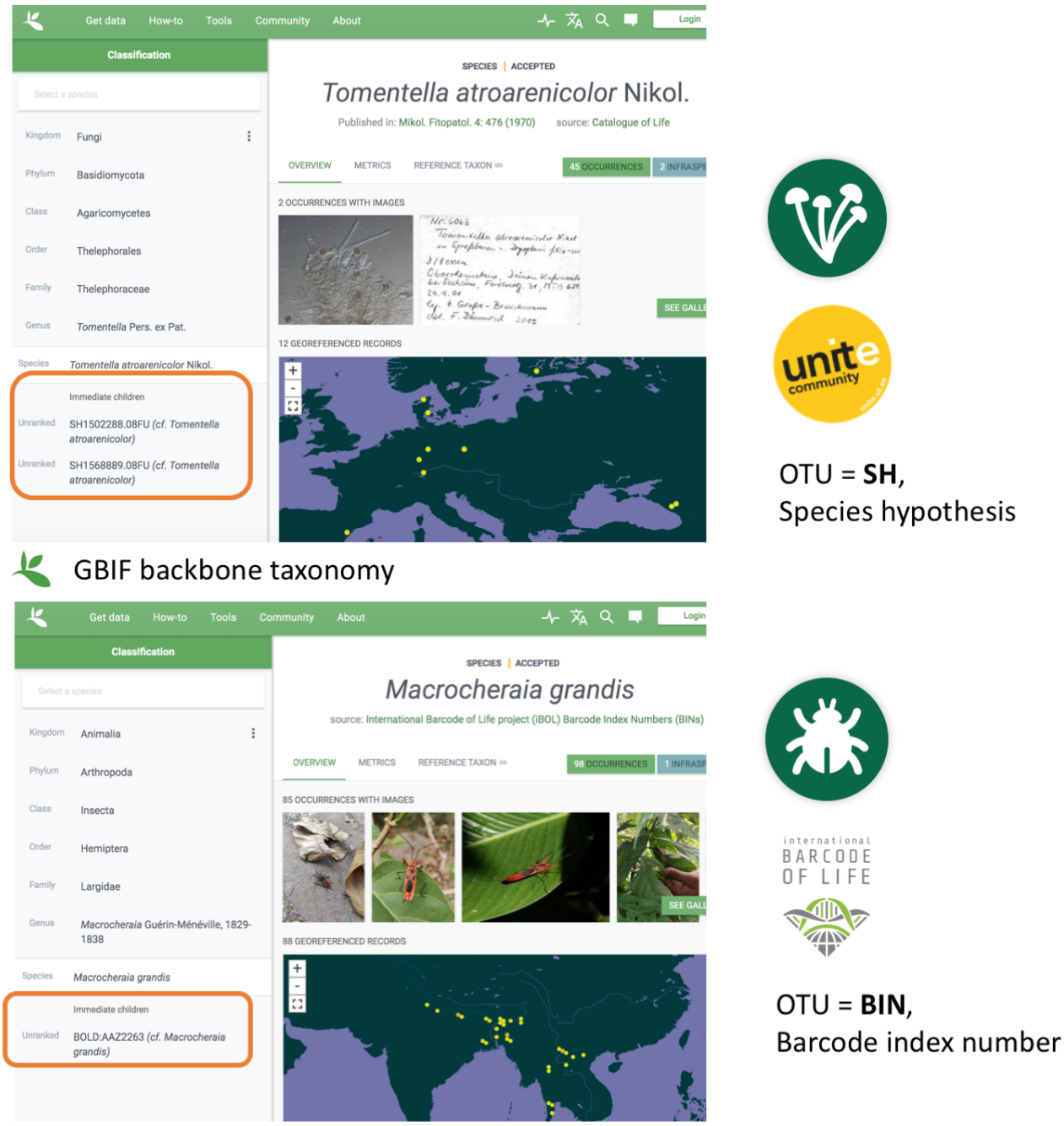

Figure 7. OTUs (SHs) from UNITE (mainly fungi, above) and from BOLD (BINs) (mainly arthropods, below) are displayed in the GBIF backbone taxonomy under their corresponding parent taxa which have Scientific names. Multiple individually observed occurrences of cryptic biodiversity become discoverable together with non-genetic evidence through a single access point.

Examples of Name references checklists

- The International Barcode of Life Consortium (2016). International Barcode of Life project (iBOL) Barcode Index Numbers (BINs). Checklist dataset https://doi.org/10.15468/wvfqoi accessed via GBIF.org on 2020-04-16.
- PlutoF (2019). UNITE - Unified system for the DNA based fungal species linked to the classification. Version 1.2. Checklist dataset https://doi.org/10.15468/mkpcy3 accessed via GBIF.org on 2020-04-16.

This guide does not provide mapping recommendations for global OTU checklists / reference libraries (Category IV), and publishing referenceable (project- or study-specific) OTU libraries as checklists is discouraged. For guidance on how to format and share OTU checklists, see the following general

Darwin Core guidelines in DwC-A template for checklists and Data quality requirements for checklists. General guidelines for MIxS checklists. For advice on how to map global reference libraries of OTUs for inclusion in the GBIF taxonomic backbone, contact the GBIF help desk.

2.1.5. Category V: Metadata-only datasets

Metadata are data about the data and is a description of the dataset in broad terms, such as authors, author affiliations, original research purpose of the dataset, DOI(s), taxonomic scope, temporal scope, and geographical scope. Information regarding laboratory methods and general sequencing methods is included in this category. This category includes datasets or collections that cannot be made available online at the moment, e.g. undigitized work.

Examples of Metadata-only datasets

- Collins E, Sweetlove M (2019). Arctic Ocean microbial metagenomes sampled aboard CGC Healy during the 2015 GEOTRACES Arctic research cruise. SCAR - Microbial Antarctic Resource System. Metadata dataset https://doi.org/10.15468/iljmun accessed via GBIF.org on 2020-04-16.
- Cary S C (2015). New Zealand Terrestrial Biocomplexity Survey. SCAR - Microbial Antarctic Resource System. Metadata dataset https://doi.org/10.15468/xnzrhq accessed via GBIF.org on 2020-04-16.

Mapping recommendations for metadata-only DNA-derived datasets (Category V) is the same as for any other metadata-only datasets, and this guide does not provide any specific mapping recommendations for metadata. Please follow general recommendations of biodiversity data portals, paying attention to required and recommended metadata. Descriptions of field, lab, and bioinformatics steps should be as detailed as possible. Describing your methods as method steps in the EML metadata makes them display on the dataset homepage in GBIF (https://www.gbif.org/dataset/3b8c5ed8-b6c2-4264-ac52-a9d772d69e9f#methodology Frøslev T, Ejrnæs R (2018). BIOWIDE eDNA Fungi dataset. Danish Biodiversity Information Facility. Occurrence dataset https://doi.org/10.15468/nesbvx accessed via GBIF.org on 2021-07-06). However, if a structured and possibly more detailed method description is already published somewhere (e.g. at protocols.io or NEON protocols collection), it is straightforward to provide a link through the MIxS SOP field (see § 2.2.1).

2.2. Data mapping

While core files store ubiquitous data on the 'what, where and when' of a record, extension files are used to describe the specifics of a certain type of observation. We propose using the DNA derived data extension to complement occurrence data derived from either barcoding, metabarcoding (eDNA) or qPCR/ddPCR. The DNA derived data extension builds on the Minimum information standards developed by the Genomic Standards Consortium (GSC) and applied by the ENA for submission of eDNA sample metadata, for example. We are following and have contributed to the guidelines proposed by the Sustainable DwC-MIxS interoperability task group under TDWG. To improve indexing and search we have opted to split some MIxS terms, for instance separating forward and reverse primer sequences and names. Furthermore, some fields from the GGBN standard and fields from the MIQE (minimum information for the publication of quantitative real-time PCR) guidelines for qPCR and ddPCR data have been included to make it applicable for a wide range of DNA-derived data.

As a first step in preparing your data for publishing, you should make sure your field names / column headers follow the Darwin Core data standard. In many cases this is straightforward, such as renaming your `lat` or `latitude` field to `decimalLatitude`. However, the Darwin Core Standard is quite flexible and some terms are used in different ways, depending on the type of data. An example of this are the fields `organismQuantity` and `organismQuantityType`, which could be used to describe the

number of individuals, per cent biomass or a score on the Braun-Blanquet Scale, as well as the number of reads of an ASV within a sample. Therefore, we here provide tables of required and recommended fields with descriptions and examples (Table 1, Table 2, Table 3 and Table 4). The recommendation to use the Occurrence core for DNA-derived data stems from the strong desire to share the sequence to help qualify the determination. Additional fields and extensions (such as extended Measurement or Fact (eMoF)) are applicable - both to occurrence cores and event core. When a sequence is derived from an organism (e.g. a parasite, gut contents, epibiont etc.) the observation may be linked to the observation of the host organism. This can be achieved using the (Resource Relation extension) of Darwin Core (e.g. https://www.gbif.org/species/143610775/verbatim). Perhaps the most important recommendation is to use globally unique (when available) and other permanent identifiers for as many data fields and parameters as possible (in all ID fields in the tables below).

2.2.1. Mapping metabarcoding (eDNA) and barcoding data

This section provides mapping recommendations for Categories I and II.

Table 2. Recommended fields for Occurrence core for Metabarcoding data

Field name	Examples	Description	Required
basisOfRecord	MaterialSample	The specific nature of the data record - a subtype of the dcterms:type. For DNA-derived occurrences, (see Category I and Category III) use MaterialSample. For enriched occurrences use PreservedSpecimen or LivingSpecimen as appropriate.	Required
occurrenceID	urn:catalog:UWBM:Bird:89776	A unique identifier for the occurrence, allowing the same occurrence to be recognized across dataset versions as well as through data downloads and use. May be a global unique identifier or an identifier specific to the data set.	Required
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	An identifier for the set of information associated with an Event (something that occurs at a place and time). May be a global unique identifier or an identifier specific to the data set.	Highly recommended
eventDate	2020-01-05	Date when the event was recorded. Recommended best practice is to use a date that conforms to ISO 8601-1:2019. For more information, check https://dwc.tdwg.org/terms/#dwc:eventDate	Required
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original Occurrence. The recommended best practice is to separate the values with a vertical bar (' | '). Including information about the observer improves the scientific reproducibility (Groom et al. 2020).	Highly recommended
organismQuantity	33	Number of reads of this OTU or ASV in the sample.	Highly recommended
organismQuantityType	DNA sequence reads	Should always be “DNA sequence reads”	Highly recommended
sampleSizeValue	1233890	Total number of reads in the sample. This is important since it allows calculating the relative abundance of each OTU or ASV within the sample. This number should preferably be calculated after universal processing (quality control, ASV denoising, chimera removal, etc.), but before manual/selective removal of e.g. non-target OTUs or ASVs from the dataset. Rarefaction (resampling to even sequencing depth across samples) is not necessary or advised.	Highly recommended

Field name	Examples	Description	Required
sampleSizeUnit	DNA sequence reads	Should always be “DNA sequence reads”	Highly recommended
materialSampleID	https://www.ncbi.nlm.nih.gov/biosample/15224856 https://www.ebi.ac.uk/ena/browser/view/SAMEA3724543 urn:uuid:a964805b-33c2-439a-beaa-6379ebbfcd03	An identifier for the MaterialSample (as opposed to a particular digital record of the material sample). Use the biosample ID if one was obtained from a nucleotide archive. In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the materialSampleID globally unique.	Highly recommended
samplingProtocol	UV light trap	The name of, reference to, or description of the method or protocol used during a sampling Event. https://dwc.tdwg.org/terms/#dwc:samplingProtocol	Recommended
associatedSequences	https://www.ncbi.nlm.nih.gov/nuccore/MK405371	A list (concatenated and separated) of identifiers (publication, global unique identifier, URI) of genetic sequence information associated with the Occurrence. Could be used for linking to archived raw barcode reads and/or associated genome sequences, e.g. in a public repository.	Recommended
identificationRemarks	RDP annotation confidence (at lowest specified taxon): 0.96, against reference database: GTDB	Specification of taxonomic identification process, ideally including data on applied algorithm and reference database, as well as on level of confidence in the resulting identification.	Recommended

Field name	Examples	Description	Required
identificationReferences	https://www.ebi.ac.uk/metagenomics/pipelines/4.1 https://github.com/terrimporter/CO1Classifier	A list (concatenated and separated) of references (publication, global unique identifier, URI) used in the Identification. Recommended best practice is to separate the values in a list with space vertical bar space (|).	Recommended
decimalLatitude	60.545207	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic centre of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.	Highly recommended
decimalLongitude	24.174556	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic centre of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.	Highly recommended
taxonID	ASV:7bdb57487bee022ba30c03c3e7ca50e1	For eDNA data, it is recommended to use an MD5 hash of the sequence and prepend it with “ASV:”. See also § 1.6.	Highly recommended, if DNA_sequence is not provided
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	Scientific name of the closest known taxon (species or higher) or an OTU identifier from BOLD (BIN) or UNITE (SH)	Required
kingdom	Animalia	Higher taxonomy	Highly recommended
phylum	Chordata	Higher taxonomy	Recommended
class	Actinopterygii	Higher taxonomy	Recommended
order	Gadiformes	Higher taxonomy	Recommended
family	Gadidae	Higher taxonomy	Recommended
genus	*Gadus*	Higher taxonomy	Recommended

Table 3. Recommended fields from the DNA derived data extension (a selection) for metabarcoding data

Field name	Examples	Description	Required
DNA_sequence	TCTATCCTCAATTATAGGTCATAATTCACCATCAGTAGATTTAGGAATTTTCTCTATTCATATTGCAGGTGTATCATCAATTATAGGATCAATTAATTTTATTGTAACAATTTTAAATATACATACAAAAACTCATTCATTAAACTTTTTACCATTATTTTCATGATCAGTTCTAGTTACAGCAATTCTCCTTTTATTATCATTA	The DNA sequence (ASV). Taxonomic interpretation of the sequence depends on the technology and reference library available at the time of publication. Hence, the most objective taxonomic handle is the sequence which can be reinterpreted in the future.	Highly recommended
sop	https://www.protocols.io/view/emp-its-illumina-amplicon-protocol-pa7dihn	Standard operating procedures used in assembly and/or annotation of genomes, metagenomes or environmental sequences. A reference to a well documented protocol, e.g. using protocols.io	Recommended
target_gene	16S rRNA, 18S rRNA, ITS	Targeted gene or marker name for marker-based studies	Highly recommended
target_subfragment	V6, V9, ITS2	Name of subfragment of a gene or markerImportant to e.g. identify special regions on marker genes like the hypervariable V6 region of the 16S rRNA gene	Highly recommended
pcr_primer_forward	GGACTACHVGGGTWTCTAAT	Forward PCR primer that was used to amplify the sequence of the targeted gene, locus or subfragment.	Highly recommended
pcr_primer_reverse	GGACTACHVGGGTWTCTAAT	Reverse PCR primer that was used to amplify the sequence of the targeted gene, locus or subfragment.	Highly recommended
pcr_primer_name_forward	jgLCO1490	Name of the forward PCR primer	Highly recommended

Field name	Examples	Description	Required
pcr_primer_name_reverse	jgHCO2198	Name of the reverse PCR primer	Highly recommended
pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	Reference for the primers	Highly recommended
env_broad_scale	forest biome [ENVO:01000174]	**Equivalent to env_biome in MIxS v4** In this field, report which major environmental system your sample or specimen came from. The systems identified should have a coarse spatial grain, to provide the general environmental context of where the sampling was done (e.g. were you in the desert or a rainforest?). We recommend using subclasses of ENVO's biome class: http://purl.obolibrary.org/obo/ENVO_00000428	Recommended
env_local_scale	litter layer [ENVO:01000338]	**Equivalent to env_feature in MIxS v4** In this field, report the entity or entities which are in your sample or specimen´s local vicinity and which you believe have significant causal influences on your sample or specimen. Please use terms that are present in ENVO and which are of smaller spatial grain than your entry for env_broad_scale.	Recommended
env_medium	soil[ENVO:00001998]	**Equivalent to env_material in MIxS v4** In this field, report which environmental material or materials (pipe separated) immediately surrounded your sample or specimen prior to sampling, using one or more subclasses of ENVO´s environmental material class: http://purl.obolibrary.org/obo/ENVO_00010483	Recommended
lib_layout	Paired	**Equivalent to lib_const_meth in MIxS v4** Specify whether to expect single, paired, or other configuration of reads	Recommended
seq_meth	Illumina HiSeq 1500	Sequencing method/platform used	Highly recommended
otu_class_appr	"dada2; 1.14.0; ASV"	Approach/algorithm and clustering level (if relevant) when defining OTUs or ASVs	Highly recommended
otu_seq_comp_appr	"blastn;2.6.0+;e-value cutoff: 0.001"	Tool and thresholds used to assign "species-level" names to OTUs or ASVs	Highly recommended

Field name	Examples	Description	Required
otu_db	"Genbank nr;221", "UNITE;8.2"	Reference database (i.e. sequences not generated as part of the current study) used to assigning taxonomy to OTUs or ASVs	Highly recommended

2.2.2. Mapping ddPCR / qPCR data

This section provides mapping recommendations for Category III.

Table 4. Recommended fields for Occurrence core for ddPCR/qPCR data

Field name	Examples	Description	Required
basisOfRecord	MaterialSample	The specific nature of the data record - a subtype of the dcterms:type. For DNA-derived occurrences (see Category I and Category III), use MaterialSample.	Required
occurrenceStatus	Present, Absent	A statement about the presence or absence of a taxon at a location.	Required
eventID	urn:uuid:a964765b-22c4-439a-jkgt-2	An identifier for the set of information associated with an Event (something that occurs at a place and time). May be a global unique identifier or an identifier specific to the dataset.	Highly recommended
eventDate	2020-01-05	Date when the event was recorded. Recommended best practice is to use a date that conforms to ISO 8601-1:2019. For more information, check https://dwc.tdwg.org/terms/#dwc:eventDate	Required
recordedBy	"Oliver P. Pearson | Anita K. Pearson"	A list (concatenated and separated) of names of people, groups, or organizations responsible for recording the original Occurrence. The recommended best practice is to separate the values with a vertical bar (' | '). Including information about the observer improves the scientific reproducibility (Groom et al. 2020).	Highly recommended
organismQuantity	50	Number of positive droplets/chambers in the sample	Highly recommended for ddPCR, dPCR
organismQuantityType	ddPCR droplets dPCR chambers	The partition type	Highly recommended for ddPCR, dPCR
sampleSizeValue	20000	The number of accepted partitions (n), e.g. meaning accepted droplets in ddPCR or chambers in dPCR.	Highly recommended for ddPCR, dPCR
sampleSizeUnit	ddPCR droplets dPCR chambers	The partition type, should be equal to the value in organismQuantityType	Highly recommended for ddPCR, dPCR

Field name	Examples	Description	Required
materialSampleID	https://www.ncbi.nlm.nih.gov/biosample/15224856 urn:uuid:a964805b-33c2-439a-beaa-6379ebbfcd03	An identifier for the MaterialSample (as opposed to a particular digital record of the material sample). Use the biosample ID if one was obtained from a nucleotide archive. In the absence of a persistent global unique identifier, construct one from a combination of identifiers in the record that will most closely make the materialSampleID globally unique.	Highly recommended
samplingProtocol	UV light trap	The name of, reference to, or description of the method or protocol used during a sampling Event. https://dwc.tdwg.org/terms/#dwc:samplingProtocol	Recommended
decimalLatitude	60.545207	The geographic latitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic centre of a Location. Positive values are north of the Equator, negative values are south of it. Legal values lie between -90 and 90, inclusive.	Highly recommended
decimalLongitude	24.174556	The geographic longitude (in decimal degrees, using the spatial reference system given in geodeticDatum) of the geographic centre of a Location. Positive values are east of the Greenwich Meridian, negative values are west of it. Legal values lie between -180 and 180, inclusive.	Highly recommended
scientificName	*Gadus morhua* L. 1758, BOLD:ACF1143	Scientific name of the closest known taxon (species or higher) or an OTU identifier from BOLD or UNITE	Required
kingdom	Animalia	Higher taxonomy	Highly recommended
phylum	Chordata	Higher taxonomy	Recommended
class	Actinopterygii	Higher taxonomy	Recommended
order	Gadiformes	Higher taxonomy	Recommended
family	Gadidae	Higher taxonomy	Recommended
genus	*Gadus*	Higher taxonomy	Recommended

Table 5. Recommended fields from the DNA derived data extension (a selection) for ddPCR/qPCR data

Field name	Examples	Description	Required
sop	https://www.protocols.io/view/protocol-for-dna-extraction-and-quantitative-pcr-d-vwie7ce https://doi.org/10.17504/protocols.io.vwie7ce	Standard operating procedures used in assembly and/or annotation of genomes, metagenomes or environmental sequences. A reference to a well documented protocol, e.g. using protocols.io	Highly recommended
annealingTemp	60	The reaction temperature during the annealing phase of PCR.	Required if annealingTemp was supplied
annealingTempUnit	Degrees Celsius		Highly recommended
pcr_cond	initial denaturation:94_3; annealing:50_1;elongation:72_1.5;final elongation:72_10;35	Description of reaction conditions and components of PCR in the form of "initial denaturation:94degC_1.5min; annealing=…"	Highly recommended
probeReporter	FAM	Type of fluorophore (reporter) used. Probe anneals within amplified target DNA. Polymerase activity degrades the probe that has annealed to the template, and the probe releases the fluorophore from it and breaks the proximity to the quencher, thus allowing fluorescence of the fluorophore.	Highly recommended
probeQuencher	NFQ-MGB	Type of quencher used. The quencher molecule quenches the fluorescence emitted by the fluorophore when excited by the cycler's light source as long as fluorophore and the quencher are in proximity, quenching inhibits any fluorescence signals.	Highly recommended
ampliconSize	83	The length of the amplicon in basepairs	Highly recommended

Field name	Examples	Description	Required
thresholdQuantificationCycle	0.3	Threshold for change in fluorescence signal between cycles	qPCR: Highly recommended
baselineValue	15	The number of cycles when fluorescence signal from the target amplification is below background fluorescence not originated from the real target amplification.	qPCR: Highly recommended
quantificationCycle	37.9450950622558	The number of cycles required for the fluorescent signal to cross a given value threshold above the baseline. Quantification cycle (Cq), threshold cycle (Ct), crossing point (Cp), and take-off point (TOP) refer to the same value from the real-time instrument. Use of quantification cycle (Cq), is preferable according to the RDML (Real-Time PCR Data Markup Language) data standard	
automaticThresholdQuantificationCycle	no	Whether the threshold was set by instrument or manually	
automaticBaselineValue	no	Whether baseline value was set by instrument or manually	
contaminationAssessment	no	Whether DNA or RNA contamination assessment was done or not	
estimatedNumberOfCopies	10300	Number of target molecules per µl. Mean copies per partition (?) can be calculated using the number of partitions (n) and the estimated copy number in the total volume of all partitions (m) with a formula ?=m/n.	
amplificationReactionVolume	22	PCR reaction volume	
amplificationReactionVolumeUnit	µl	Unit used for PCR reaction volume. Many of the instruments require preparation of a much larger initial sample volume than is actually analysed.	
pcr_analysis_software	BIO-RAD QuantaSoft	The program used to analyse the d(d)PCR runs.	
experimentalVariance		Multiple biological replicates are encouraged to assess total experimental variation. When single dPCR experiments are performed, a minimal estimate of variance due to counting error alone must be calculated from the binomial (or suitable equivalent) distribution.	

Field name	Examples	Description	Required
target_gene	16S rRNA, 18S rRNA, nif, amoA, rpo	Targeted gene or marker name for marker-based studies	Highly recommended
target_subfragment	V6, V9, ITS	Name of subfragment of a gene or markerImportant to identify, for example, special regions on marker genes like the hypervariable V6 region of the 16S rRNA gene	Highly recommended
pcr_primer_forward	GGACTACHVGGGTWTCTAAT	Forward PCR primer that was used to amplify the sequence of the targeted gene, locus or subfragment.	Highly recommended
pcr_primer_reverse	GGACTACHVGGGTWTCTAAT	Reverse PCR primer that was used to amplify the sequence of the targeted gene, locus or subfragment.	Highly recommended
pcr_primer_name_forward	jgLCO1490	Name of the forward PCR primer	Highly recommended
pcr_primer_name_reverse	jgHCO2198	Name of the reverse PCR primer	Highly recommended
pcr_primer_reference	https://doi.org/10.1186/1742-9994-10-34	Reference for the primers	Highly recommended
env_broad_scale	forest biome [ENVO:01000174]	**Equivalent to env_biome in MIxS v4** In this field, report which major environmental system your sample or specimen came from. The systems identified should have a coarse spatial grain, to provide the general environmental context of where the sampling was done (e.g. were you in the desert or a rainforest?). We recommend using subclasses of ENVO´s biome class: http://purl.obolibrary.org/obo/ENVO_00000428	Recommended
env_local_scale	litter layer [ENVO:01000338]	**Equivalent to env_feature in MIxS v4** In this field, report the entity or entities which are in your sample or specimen´s local vicinity and which you believe have significant causal influences on your sample or specimen. Please use terms that are present in ENVO and which are of smaller spatial grain than your entry for env_broad_scale.	Recommended

Field name	Examples	Description	Required
env_medium	soil [ENVO:00001998]	**Equivalent to env_material in MIxS v4** In this field, report which environmental material or materials (pipe separated) immediately surrounded your sample or specimen prior to sampling, using one or more subclasses of ENVO´s environmental material class: http://purl.obolibrary.org/obo/ENVO_00010483	Recommended
concentration	67.5	Concentration of DNA (weight ng/volume µl) see also http://terms.tdwg.org/wiki/ggbn:concentration	Recommended
concentrationUnit	ng/µl	Unit used for concentration measurement see also http://terms.tdwg.org/wiki/ggbn:concentrationUnit	Recommended
methodDeterminationConcentrationAndRatios	Nanodrop, Qubit	Description of method used for concentration measurement see also http://terms.tdwg.org/wiki/ggbn:methodDeterminationConcentrationAndRatios	Recommended
ratioOfAbsorbance260_230	1.89	Ratio of absorbance at 260 nm and 230 nm assessing DNA purity (mostly secondary measure, indicates mainly EDTA, carbohydrates, phenol), (DNA samples only). see also http://terms.tdwg.org/wiki/ggbn:ratioOfAbsorbance260_230	Recommended
ratioOfAbsorbance260_280	1.91	Ratio of absorbance at 280 nm and 230 nm assessing DNA purity (mostly secondary measure, indicates mainly EDTA, carbohydrates, phenol), (DNA samples only). see also http://terms.tdwg.org/wiki/ggbn:ratioOfAbsorbance260_280	Recommended
samp_collect_device	biopsy, niskin bottle, push core	The method or device employed for collecting the sample	Recommended
samp_mat_process	filtering of seawater, storing samples in ethanol	Any processing applied to the sample during or after retrieving the sample from environment. This field accepts OBI, for a browser of OBI (v 2018-02-12) terms please see http://purl.bioontology.org/ontology/OBI	Recommended
samp_size	5 litre	Amount or size of sample (volume, mass or area) that was collected	Recommended
size_frac	0-0.22 micrometer	Filtering pore size used in sample preparation	Recommended
pcr_primer_lod	51	The assay's ability to detect the target at low levels	Highly recommended
pcr_primer_loq	184	The assay's ability to quantify copy number at low levels	Highly recommended

2.3. Marine datasets and the Ocean Biodiversity Information System (OBIS)

When working with datasets originating in the marine environment, it is recommended that the information is published also in the Ocean Biodiversity Information System (OBIS) in addition to GBIF. OBIS is a global biodiversity database, which is specialized in providing reliable and accessible data related to marine life and is a part of the IOC-UNESCO. Like GBIF, and ALA, OBIS uses the DwC-A format for data indexing and publishing. By publishing marine datasets through OBIS in addition to other biodiversity databases, the data can reach a broader audience, and diverse groups working in the field of marine biodiversity, as datasets in OBIS are often used for UN processes. With the focus on marine datasets, stringent quality controls on the data increase the reliability of the data and lead to small differences in what information is required for publishing in OBIS as opposed to GBIF.

To ensure consistent taxonomic nomenclature OBIS uses the World Register of Marine Species (WoRMS) as the only taxonomic backbone. This is the case also for occurrences derived from genetic data; a scientific name linked to a scientific name ID from the WoRMS database is highly recommended information for publishing. If a scientific name ID is not provided, OBIS will try to match the scientific name with WoRMS during ingestion, but this should be avoided whenever possible. Scientific names not listed in WoRMS are acceptable, and will be submitted to WoRMS for review and possible inclusion in the register. Fully unclassified sequences are recommended to be classified as "incertae sedis", with the WoRMS `scientificNameID` urn:lsid:marinespecies.org:taxname:12. This will ensure correct interpretation by both GBIF and OBIS. Additionally, it is recommended that sequence identifiers from the used reference databases (e.g. Barcode index numbers: BINs from BOLD) be added in the `taxonConceptID` field of the occurrence core table. In this way OBIS will retain its taxonomic backbone based on WoRMS, while enabling linking to disparate reference sequence databases. Names from reference databases which are not strictly scientific names, can be added as `verbatimIdentification`. Automatic classification of species names can often be done through the WoRMS taxon match services and R packages like worrms and taxize. In the future, OBIS plans to periodically search and update the taxonomic assignments of submitted sequences as reference databases develop with time, so recording the sequence information linked to each occurrence is highly recommended.

Another required field in OBIS data submissions are geographic coordinates. OBIS performs additional quality checks related to marine data; e.g. that coordinates for strictly marine species are not on land, and that the depth value reported is in a reasonable range. Finally, it should be mentioned that in addition OBIS supports the use of the extended Measurement or Fact (eMoF). This extension allows linking environmental data and sampling facts to sampling events or occurrences, as well as biological measurements to occurrences in a flexible and standardized manner. OBIS has an example eDNA metabarcoding dataset with scripts for data formatting available at https://github.com/iobis/dataset-edna.

Table 6. OBIS requirements and recommendations for recording DNA-based occurrences. The table highlights important differences in field values and requirements compared to when publishing to GBIF. Here exemplified with a DNA detection of Blue whale (Balaenoptera musculus).

Field name	Value/example (OBIS)	Description	Required
scientificName	Balaenoptera musculus	Scientific name, preferably as listed in the WoRMS database. This differs from GBIF, where it is recommended to use the taxon name derived from the classification approach used.	Required
scientificNameID	urn:lsid:marinespecies.org:taxname:137090	The scientific name ID of "Balaenoptera musculus" as per the WoRMS database.	Highly recommended
taxonConceptID	NCBI:txid9771	The NCBI ID linked to Balaenoptera musculus in the NCBI taxonomic database. Can also be a BIN-ID if BOLD was used for identification, or another ID from a different database.	Recommended
verbatimIdentification	Balaenoptera musculus	The name corresponding to the NCBI ID (Balaenoptera musculus) (or other ID). This does not necessarily correspond to the value in scientific name.	Recommended

Table 7. OBIS requirements and recommendations for recording sequences that cannot be classified to a scientific name at any taxonomic level.

Field name	Value (OBIS)	Description	Required
scientificName	incertae sedis	The scientific name for unknown sequences recommended by OBIS. Use this name when the sequence/taxonomy is unknown. This differs from GBIF, where it is recommended to use the taxon name as retrieved from the classifier even when it is not strictly a scientific name.	Required
scientificNameID	urn:lsid:marinespecies.org:taxname:12	The scientific name ID of "incertae sedis" as per the WoRMS database for unknown sequences recommended by OBIS. Use this ID when the sequence/taxonomy is unknown.	Highly recommended
taxonConceptID	NCBI:txid1899546	The ID in an external taxonomic database, like a sequence reference database for example.	Recommended
verbatimIdentification	Phototrophic eukaryote	The name of the taxon in an external database, corresponding to the taxon concept ID.	Recommended

3. Future prospects

The present interest in exposing DNA-derived data through biodiversity data platforms is very high, and it is very likely that the demand will grow. Our aim is for the mapping recommendations provided here to remain valid and evolve slowly, even as packaging and indexing by biodiversity data platforms may develop more rapidly. The authors are aware of but did not yet consult the BOLD Handbook, BIOM format and http://edamontology.org/page.

We suggest that data platforms such as ALA and GBIF work towards adopting data formats that support more complex relational and hierarchical data. Examples could be the Frictionless Data Format and the more domain-specific Biological Observation Matrix (BIOM) format. The latter is used by several bioinformatic tools (QIIME2, Mothur, USEARCH etc.), and hence could help publishers skip a step in converting data into DwC-A format. A more flexible data format than the current DwC star schema is crucial for allowing hierarchical sampling events and material samples as well as attaching sequence data to individual occurrences within a sampling event.

Biodiversity data platforms will also need to enable researchers to easily include or exclude DNA-derived occurrence data from their query results. The data formats suggested above could open opportunities for a richer classification of the types of evidence on which a specific occurrence record is based. However, for the time being there is a lack of an appropriate value in the BasisOfRecord vocabulary for these data types. We suggest, as a pragmatic immediate solution, that the BasisOfRecord is extended with a value such as “DNA”, “DNA-derived”, or similar. As described above, DNA-derived data may come from well-documented sampling or individual organisms, may be backed by preserved physical material or not, and may result from genetic sequencing or other DNA detection methods, such as qPCR. Biodiversity data platforms and TDWG should provide the means of differentiating between these data types and their origins.

We also recommend that the data platforms index the actual sequences, or at least a MD5 checksum of these, to facilitate searches for ASVs across datasets. If ASVs are provided, MD5s should be generated by the biodiversity discovery platforms; if ASVs are not provided, MD5s need to be mandatory.

As mentioned in § 1.6 and § 2.1.4, we encourage the biodiversity data platforms to continue work on adopting relevant molecular taxonomic reference databases into their taxonomic backbones.

Broader application of other methods and technologies, such as Oxford Nanopore, PacBio and shotgun sequencing, will likely trigger the need for adjustments to this guide to accommodate specific new data and metadata fields.

Glossary

Atlas of Living Australia (ALA)

The ALA is a web-based platform that pulls together Australian biodiversity data from multiple sources, making it accessible and reusable to anyone (see https://www.ala.org.au/about-ala/). The open infrastructure platform developed by the ALA is also used by several other countries for their own national biodiversity data platform (see https://living-atlases.gbif.org/).

Amplicon Sequence Variant (ASV)

Unique DNA sequence derived from high-throughput sequencing and denoising, and assumed to represent a biologically real sequence variant. See also Operational Taxonomic Unit (OTU) and (Callahan et al. 2017).

Application Programming Interface (API)

Set of protocols and tools for interaction and data transmission between different computer applications.

Barcode Index Numbers (BINs)

Species-level Operational Taxonomic Units (OTUs) derived from clustering of the cytochrome c oxidase I (COI) gene in animals. Each BIN is assigned a globally unique identifier, and is made available in searchable database within the Barcode of Life Data System (BOLD).

Barcode of Life Data System (BOLD)

BOLD is the reference database maintained by the Centre for Biodiversity Genomics in Guelph on behalf of the International Barcode of Life Consortium (IBOL). It hosts data on barcode reference specimens and sequences for eukaryote species, particularly COI for animals, and maintains the Barcode Index Number (BIN; Ratnasingham & Hebert 2013) system, identifiers for OTUs of approximately species rank, based on clusters of closely similar sequences.

Biodiversity data platform

General online resource to discover and access biodiversity data derived from various sources, such as natural history collections, citizen science, ecology and monitoring projects, and genetic sequences. Can be global (GBIF) or national (ALA).

Clustering

In taxonomic classification, the process of grouping organisms together according to some similarity criterion. See Operational Taxonomic Unit.

Community (bulk) DNA

DNA from bulk samples (e.g. plankton samples or Malaise trap samples consisting of several individuals from many species). For the purpose of this guide, bulk sample DNA is included in the eDNA concept.

Darwin Core Archive (DwC-A)

Compressed (ZIP) file format for exchange of biodiversity data compiled in accordance with the Darwin Core (DwC) standard. Essentially a self-contained set of interconnected CSV files and an XML document describing included files and data columns, and their mutual relationships.

Darwin Core (DwC) standard

Standard for sharing and publishing biodiversity data, originating from the Biodiversity Information Standards (TDWG) community. In principle, a set of terms used for describing different entities of biodiversity observations, such as sampling events, occurrences and taxa. Current Darwin Core terms are described in the Quick Reference Guide.

Data vocabulary

Set of preferred terms or concepts with specific, well-defined meanings and interrelationships, facilitating data exchange and reuse.

ddPCR (droplet digital Polymerase Chain Reaction)

Droplet digital PCR. Method for measuring absolute amount of DNA (number of copies) of one marker in a sample. See also qPCR.

Denoising

In metabarcoding, method for separation of true biological sequences (see ASVs) from spurious sequence variants caused by PCR amplification and sequencing error.

Digital Object Identifier (DOI)

Long-lasting reference used to uniquely identify (and locate) digital information objects, such as a biodiversity data set or a scientific publication.

DNA barcoding and metabarcoding (amplicon sequencing)

Use of short, standardized DNA fragments to identify individual organisms via sequencing. Metabarcoding combines barcoding with high-throughput DNA sequencing, using universal primers to amplify and sequence large groups of organisms in eDNA samples.

DNA marker

A DNA fragment used as a marker of some property (e.g., taxonomic affiliation). May, but does not have to, be a gene or a part of a gene.

DNA metabarcoding database

Database containing DNA sequences (DNA barcodes) from previously recovered or studied organisms. The reference sequences were ideally generated from individuals of described, well-studied species-with the type specimen serving as the ideal-or higher taxonomic level (e.g., genus, family), but may also stem from eDNA sequencing efforts. It is wise not to trust "reference sequences" blindly.

DNA probe

A short, synthetic single-stranded DNA fragment with fluorescent labelling that binds to a selected region of target DNA (marker) during PCR. Increases specificity and can be used in addition to primers in qPCR and ddPCR to detect and quantify a genetic marker.

European Bioinformatics Institute (EMBL-EBI)

Intergovernmental organization for bioinformatics research and services, part of the European Molecular Biology Laboratory (EMBL), providing eg. (raw) sequence reads and assembly data via the European Nucleotide Archive (ENA).

Environmental DNA (eDNA)

DNA from an environmental sample, e.g. soil, water, air or host organism. An often used definition is that environmental DNA is the genetic material (DNA) obtained from environmental samples without any obvious evidence of biological source material (Thomsen and Willerslev 2015).

European Nucleotide Archive (ENA)

European repository for nucleotide sequences, covering raw sequencing data, sequence assembly information and functional annotation. Includes the Sequence Read Archive (SRA), and is maintained by the European Bioinformatics Institute (EMBL-EBI), as part of the International Nucleotide Sequence Database Collaboration (INSDC).

FASTQ

Text-based standard for storing molecular sequences and associated quality measures deriving from High-throughput sequencing (HTS). For each sequence position, single ASCII-characters are used to represent base call (identified nucleotide) and score, respectively.

Global Biodiversity Information Facility (GBIF)

International network and research infrastructure, mainly focused on mobilizing and providing open access to global biodiversity data.

Global Genome Biodiversity Network (GGBN)

International network of institutions concerned with efficient sharing and usage of genomic biodiversity samples and associated metadata, e.g. promoting the Darwin Core-compatible GGBN

Data Standard.

Global Positioning System (GPS)

Satellite navigation system operated by the United States Space Force.

High-throughput sequencing (HTS)

Different technologies for massively parallel sequencing, producing millions of DNA sequence reads from library preparations of genetic material, rather than targeting single amplicons as in traditional Sanger sequencing. Also called Next Generation Sequencing (NGS).

Ingestion

Process of importing data from heterogeneous sources, such as local databases, text files or spreadsheets, to a common destination system, such as an online biodiversity data platform, for storage and further analysis. Typically includes steps of extraction, transformation (cleaning) and loading (ETL).

Indexing

Organization of information in accordance with a specific schema or structure, making data easier to access and present.

International Nucleotide Sequence Database Collaboration (INSDC)

Joint effort of the DNA Databank of Japan (DDBJ), EMBL and NCBI to provide global public access to nucleotide sequence data and associated information.

Metagenomics

PCR-free sequencing of random genomic fragments in a mixed sample.

Minimum Information about any (x) Sequence (MIxS) standard

Family of standards (checklists) for sequence metadata, developed by the Genomic Standards Consortium (GSC).

molecular Operational Taxonomic Unit (mOTU)

See Operational Taxonomic Unit (OTU).

National Center for Biotechnology Information (NCBI)

Division of United States National Library of Medicine (NLM) housing important bioinformatics resources, such as the GenBank database of DNA sequences, and the Sequence Read Archive (SRA) of high throughput sequencing data.

Next Generation Sequencing (NGS)

See High-throughput sequencing (HTS).

Occurrence

An existence of an Organism (sensu http://rs.tdwg.org/dwc/terms/Organism) at a particular place at a particular time.

Operational Taxonomic Unit (OTU)

Cluster of organisms based on similarity in specific DNA marker sequence(s), used for taxonomic classification. Includes, for example, Species Hypothesis in UNITE, and Barcode Index Numbers in the Barcode of Life Data System (BOLD). Amplicon Sequence Variants (ASVs) may be considered analogous to zero radius OTUs (zOTUs).

Polymerase Chain Reaction (PCR)

Technique for fast amplification and detection of specific fragments of target DNA (or RNA) sequences. Amplified regions are determined by the pair of PCR primers used in the reaction.

Pipeline

In bioinformatics, a set of algorithms or tools applied in a predefined workflow to process e.g. High-throughput sequencing (HTS) data.

Primers (PCR primers)

Short, synthetic, single-stranded DNA fragments that bind to a selected region of target DNA (marker) to initiate replication during PCR. A pair of primers is necessary for the polymerase enzyme to amplify the selected marker.

qPCR (quantitative Polymerase Chain Reaction)

Quantitative PCR. Method that measures relative DNA quantity of a marker in a sample. See also ddPCR.

Sample

Material (water, soil, gut content, etc) obtained for analysis.

Sequence alignment

Bioinformatic process of comparing and arranging two or more molecular (DNA, RNA or protein) sequences to detect similarities caused by e.g. evolutionary relatedness.

Species Hypothesis (SH)

Species-level Operational Taxonomic Unit (OTU) as defined in the UNITE database and sequence management environment, for Fungi.

Specimen

An individual animal, plant, fungus, etc. used as an example of its species or type for scientific study or display.

Sequence Read Archive (SRA)

Public repository of high throughput (NGS) sequencing data, with instances operated by the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EMBL-EBI), and the DNA Data Bank of Japan (DDBJ). Includes both raw (non-denoised) sequencing output and sequence alignments. One of three components of the European Nucleotide Archive (ENA), and previously known as the Short Read Archive.

Target-capture sequencing

Sequencing of DNA fragments isolated with hybridization probes.

UNITE

UNITE is a web-based sequence management environment centred on the eukaryotic nuclear ribosomal ITS region. All public sequences are clustered into species hypotheses (SHs), which are assigned unique DOIs. An SH-matching service outputs various elements of information, including what species are present in eDNA samples, whether these species are potentially undescribed new species, other studies in which they were recovered, whether the species are alien to a region, and whether they are threatened. The DOIs are connected to the taxonomic backbone of the PlutoF platform and GBIF, such that they are accompanied by a taxon name where available. The data used in UNITE are hosted and managed in PlutoF. Data are represented through a range of standards, primarily Darwin Core, MIxS, and DMP Common Standard; partial support is available for EML, MCL, and GGBN. PlutoF exports data primarily through the CSV and FASTA formats. PlutoF

can also be used to publish data in GBIF (using the DwC format) and to prepare GenBank submission files. It is furthermore possible to download species lists from your data and download your project as a JSON document with project data in hierarchically structured.

Zero radius otu (zOTU)

See ASV.

References

- Amid C, Alako BT, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J, Fan J, Harrison PW, Holt S, Hussein A, Ivanov E & Jayathilaka S (2020) The European Nucleotide Archive in 2019. Nucleic acids research 48(D1): D70-D76. https://doi.org/10.1093/nar/gkz1063
- Andersen K, Bird KL, Rasmussen M, Haile J, Breuning-Madsen H, Kjaer KH, Orlando L, Gilbert MTP and Willerslev E (2012) Meta-Barcoding of 'Dirt' DNA from Soil Reflects Vertebrate Biodiversity. Molecular Ecology 21(8): 1966–79. https://doi.org/10.1111/j.1365-294X.2011.05261.x
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank, Nucleic Acids Research, 34(1): D16–D20, https://doi.org/10.1093/nar/gkj157
- Berry O, Jarman S, Bissett A, Hope M, Paeper C, Bessey C, Schwartz MK, Hale J & Bunce M (2021) Making environmental DNA (eDNA) biodiversity records globally accessible. Environmental DNA, 3(4), 699-705. https://doi.org/10.1002/edn3.173
- Bessey C, Jarman SN, Berry O et al. (2020) Maximizing fish detection with eDNA metabarcoding. Environmental DNA: 1–12. https://doi.org/10.1002/edn3.74
- Biggs J, Ewald N, Valentini A, Gaboriaud C, Dejean T, Griffiths RA, Foster J, et al. (2015) Using eDNA to Develop a National Citizen Science-Based Monitoring Programme for the Great Crested Newt (*Triturus cristatus*). Biological Conservation 183: 19–28. https://doi.org/10.1016/j.biocon.2014.11.029
- Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R & Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society B: Biological Sciences 360(1462): 1935-1943. https://doi.org/10.1098/rstb.2005.1725
- Bolyen E, Rideout JR, Dillon MR et al. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9
- Boussarie G, Bakker J, Wangensteen OS, Mariani S, Bonnin L, Juhel JB, Kiszka JJ, Kulbicki M, Manel S, Robbins WD & Vigliola L (2018) Environmental DNA illuminates the dark diversity of sharks. Science Advances 4(5): eaap9661. https://doi.org/10.1126/sciadv.aap9661
- Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, … & Wittwer CT (2009). The MIQE Guidelines: M inimum I nformation for Publication of Q uantitative Real-Time PCR Experiments. https://doi.org/10.1373/clinchem.2008.112797
- Callahan B, McMurdie P & Holmes S (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal 11: 2639–2643. https://doi.org/10.1038/ismej.2017.119
- Callahan B, McMurdie P, Rosen M et al. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods 13, 581–583. https://doi.org/10.1038/nmeth.3869
- Centre for Biodiversity Genomics, University of Guelph (2021) The Global Taxonomy Initiative 2020: A Step-by-Step Guide for DNA Barcoding. Technical Series No. 94. Secretariat of the Convention on Biological Diversity, Montreal, 66 pp. https://www.cbd.int/doc/publications/cbd-ts-94-en.pdf
- Convention on Biological Diversity (2020) Report of the ad hoc Technical Expert Group on Digital Sequence Information On Genetic Resources, 17-20 March 2020. Montreal, Canada. https://www.cbd.int/doc/c/ba60/7272/3260b5e396821d42bc21035a/dsi-ahteg-2020-01-07-en.pdf
- Debroas D, Domaizon I, Humbert JF, Jardillier L, Lepère C, Oudart A & Taïb N (2017) Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data. FEMS Microbiology Ecology 93(4): fix023. https://doi.org/10.1093/femsec/fix023

- Doi H, Fukaya K, Oka SI, Sato K, Kondoh M & Miya M (2019) Evaluation of Detection Probabilities at the Water-Filtering and Initial PCR Steps in Environmental DNA Metabarcoding Using a Multispecies Site Occupancy Model. Scientific Reports 9(1): 3581. https://doi.org/10.1038/s41598-019-40233-1
- Durkin L, Jansson T, Sanchez M, Khomich M, Ryberg M, Kristiansson E, Nilsson RH (2020) When mycologists describe new species, not all relevant information is provided (clearly enough). MycoKeys 72: 109-128. https://doi.org/10.3897/mycokeys.72.56691
- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST, Bioinformatics 26(19): 2460–2461, https://doi.org/10.1093/bioinformatics/btq461
- Ekrem T & Majaneva M (2019) DNA-Metastrekkoding Til Undersøkelser Av Invertebrater I Ferskvann. NTNU Vitenskapsmuseet Naturhistorisk Notat. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2612638.
- Elbrecht V & Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. PLoS ONE 10(7): e0130324. https://doi.org/10.1371/journal.pone.0130324
- Ficetola GF, Miaud C, Pompanon F, & Taberlet P (2008). Species detection using environmental DNA from water samples. Biology letters, 4(4), 423-425. https://doi.org/10.1098/rsbl.2008.0118
- Fossøy F, Brandsegg H, Sivertsgård R, Pettersen O, Sandercock BK, Solem Ø, Hindar K & Tor AM (2019) Monitoring Presence and Abundance of Two Gyrodactylid Ectoparasites and Their Salmonid Hosts Using Environmental DNA. Environmental DNA. https://doi.org/10.1002/edn3.45.
- Frøslev TG, Kjøller R, Bruun HH et al. (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nat Commun 8, 1188 . https://doi.org/10.1038/s41467-017-01312-x
- Groom Q, Güntsch A, Huybrechts P, Kearney N, Leachman S, Nicolson N, Page RDM, Shorthouse DP, Thessen, AE, Haston E. People are essential to linking biodiversity data. 2020. Database 2020:baaa072 https://doi.org/10.1093/database/baaa072.
- Hernandez C, Bougas B, Perreault-Payette A, Simard A, Côté G, & Bernatchez L (2020). 60 specific eDNA qPCR assays to detect invasive, threatened, and exploited freshwater vertebrates and invertebrates in Eastern Canada. Environmental DNA, 2(3): 373-386. https://doi.org/10.1002/edn3.89
- Hofstetter, V, Buyck, B, Eyssartier, G, Schnee S, Gindro K (2019) The unbearable lightness of sequenced-based identification. Fungal Diversity 96, 243–284. https://doi.org/10.1007/s13225-019-00428-3
- Huggett JF, Foy CA, Benes V, Emslie K, Garson JA, Haynes R, … & Bustin SA (2013). The Digital MIQE Guidelines: M inimum I nformation for Publication of Q uantitative Digital PCR E xperiments. Clinical chemistry, 59(6), 892-902. https://doi.org/10.1373/clinchem.2013.206375
- Hugerth LW, Andersson AF (2017) Analysing Microbial Community Composition through Amplicon Sequencing: From Sampling to Hypothesis Testing. Frontiers in Microbiology 8: 1561. https://doi.org/10.3389/fmicb.2017.01561
- Knudsen SW, Ebert RB, Hesselsøe M, Kuntke F, Hassingboe J, Mortensen PB, Thomsen PF et al (2019) Species-Specific Detection and Quantification of Environmental DNA from Marine Fishes in the Baltic Sea. Journal of Experimental Marine Biology and Ecology 510: 31–45. https://doi.org/10.1016/j.jembe.2018.09.004
- Lacoursière-Roussel A, Rosabal M & Bernatchez L (2016) Estimating Fish Abundance and Biomass from eDNA Concentrations: Variability among Capture Methods and Environmental Conditions. Molecular Ecology Resources 16(6): 1401–14. https://doi.org/10.1111/1755-0998.12522

- Leebens-Mack J, Vision T, Brenner E, Bowers JE, Cannon S, Clement MJ, Cunningham CW, DePamphilis C, DeSalle R, Doyle JJ & Eisen JA (2006) Taking the first steps towards a standard for reporting on phylogenies: Minimum Information About a Phylogenetic Analysis (MIAPA). Omics: a journal of integrative biology 10(2): 231-237. https://doi.org/10.1089/omi.2006.10.231
- Leinonen R, Sugawara H, Shumway M & International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. Nucleic Acids Research 39(suppl_1): D19-D21. https://doi.org/10.1093/nar/gkq1019
- Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. PeerJ 2:e593 https://doi.org/10.7717/peerj.593
- McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, … & Caporaso JG (2012). The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience, 1(1), 2047-217X. https://doi.org/10.1186/2047-217X-1-7
- Miralles A, Bruy T, Wolcott K, Scherz MD, Begerow D, Beszteri B, Bonkowski M, Felden J, Gemeinholzer B, Glaw F & Glöckner FO (2020) Repositories for Taxonomic Data: Where We Are and What is Missing. Systematic Biology: syaa026. https://doi.org/10.1093/sysbio/syaa026
- Mora C, Tittensor DP, Adl S, Simpson AG & Worm B (2011) How many species are there on Earth and in the ocean? PLoS Biology 9(8): e1001127. https://doi.org/10.1371/journal.pbio.1001127
- Nilsson RH, Tedersoo L, Abarenkov K, Ryberg M, Kristiansson E, Hartmann M, Schoch CL, Nylander JA, Bergsten J, Porter TM & Jumpponen A (2012) Five simple guidelines for establishing basic authenticity and reliability of newly generated fungal ITS sequences. MycoKeys 4: 37-63. https://doi.org/10.3897/mycokeys.4.3606
- Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, 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, Volume 47, Issue D1, D259–D264. https://doi.org/10.1093/nar/gky1022
- Ogram A, Sayler GS, Barkay T (1987) The Extraction and Purification of Microbial DNA from Sediments. Journal of Microbiological Methods. https://doi.org/10.1016/0167-7012(87)90025-x.
- Ovaskainen O, Schigel D, Ali-Kovero H et al. (2013) Combining high-throughput sequencing with fruit body surveys reveals contrasting life-history strategies in fungi. The ISME Journal 7: 1696–1709. https://doi.org/10.1038/ismej.2013.61
- Parks, DH, Chuvochina, M, Chaumeil, P, Rinke C, Mussig AJ, Hugenholtz P (2020) A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol 38, 1079–1086. https://doi.org/10.1038/s41587-020-0501-8
- Pearson, WR & Lipman DJ (1988) Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences 85(8): 2444-2448. https://dx.doi.org/10.1073%2Fpnas.85.8.2444
- Penev P, Mietchen D, Chavan VS, Hagedorn G, Smith VS, Shotton D, Tuama ÉÓ, Senderov V, Georgiev T, Stoev P, Groom QJ, Remsen D, Edmunds SC (2017) Strategies and guidelines for scholarly publishing of biodiversity data. Research ideas and outcomes 3: e12431, https://doi.org/10.3897/rio.3.e12431
- Pietramellara G, Ascher J, Borgogni F, Ceccherini MT, Guerri G & Nannipieri P (2009) Extracellular DNA in Soil and Sediment: Fate and Ecological Relevance. Biology and Fertility of Soils 45: 219-235. https://doi.org/10.1007/s00374-008-0345-8.
- Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System. Molecular Ecology Notes, 7: 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
- Ratnasingham S, Hebert PDN (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PloS one, 8(7), e66213. https://doi.org/10.1371/journal.pone.0066213

- Rognes T, Flouri T, Nichols B, Quince C, Mahé F (2016) VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584. https://doi.org/10.7717/peerj.2584
- Ruppert KM, Kline RJ, Rahman MS (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Global Ecology and Conservation, 17, e00547. https://doi.org/10.1016/j.gecco.2019.e00547
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, … & Weber CF (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology, 75(23), 7537-7541. https://doi.org/10.1128/AEM.01541-09
- Shea MM, Kuppermann J, Rogers MP, Smith DS, Edwards P & Boehm AB (2023) Systematic review of marine environmental DNA metabarcoding studies: toward best practices for data usability and accessibility. PeerJ, 11, p.e14993. https://doi.org/10.7717/peerj.14993
- Sigsgaard EE, Jensen MR, Winkelmann IE, Møller PR, Hansen MM, Thomsen PF (2020). Population -level inferences from environmental DNA—Current status and future perspectives. Evolutionary Applications, 13(2), 245-262. https://doi.org/10.1111/eva.12882
- Somervuo P, Koskela S, Pennanen J, Nilsson RH, Ovaskainen O (2016) Unbiased probabilistic taxonomic classification for DNA barcoding. Bioinformatics 32(19):2920–2927, https://doi.org/10.1093/bioinformatics/btw346
- Strand DA, Johnsen SI, Rusch JC, Agersnap S, Larsen WB, Knudsen SW, Møller PR & Vrålstad T (2019) Monitoring a Norwegian Freshwater Crayfish Tragedy: eDNA Snapshots of Invasion, Infection and Extinction. Journal of Applied Ecology 56(7): 1661-1673. https://doi.org/10.1111/1365-2664.13404.
- Taberlet P, Bonin A, Coissac E & Zinger L (2018) Environmental DNA: For Biodiversity Research and Monitoring. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780198767220.001.0001
- Taberlet P, Coissac E, Hajibabaei M & Rieseberg LH (2012) Environmental DNA. Molecular Ecology 21(8): 1789–93. https://doi.org/10.1111/j.1365-294X.2012.05542.x
- Takahara T, Minamoto T, Yamanaka H, Doi H & Kawabata Z (2012) Estimation of Fish Biomass Using Environmental DNA. PLoS ONE 7(4): e35868. https://doi.org/10.1371/journal.pone.0035868
- Tedersoo, L, Bahram M, Puusepp R, Nilsson RH & James TY (2017) Novel soil-inhabiting clades fill gaps in the fungal tree of life. Microbiome 5: 42. https://doi.org/10.1186/s40168-017-0259-5
- Tedesco PA, Bigorne R, Bogan AE, Giam X, Jézéquel C & Hugueny B (2014) Estimating how many undescribed species have gone extinct. Conservation Biology 28(5): 1360-1370. https://doi.org/10.1111/cobi.12285
- Thalinger B, Deiner K, Harper LR, Rees HC, Blackman RC, Sint D, … & Bruce K (2021). A validation scale to determine the readiness of environmental DNA assays for routine species monitoring. Environmental DNA. https://doi.org/10.1101/2020.04.27.063990
- Thomsen PF, Kielgast JOS, Iversen LL, Wiuf C, Rasmussen M, Gilbert MTP Orlando L & Willerslev E (2012) Monitoring Endangered Freshwater Biodiversity Using Environmental DNA. Molecular Ecology 21(11): 2565–73. https://doi.org/10.1111/j.1365-294X.2011.05418.x
- Thomsen PF, Møller PR, Sigsgaard EE, Knudsen SW, Jørgensen OA & Willerslev E (2016) Environmental DNA from Seawater Samples Correlate with Trawl Catches of Subarctic, Deepwater Fishes. PLoS ONE 11(11): e0165252. https://doi.org/10.1371/journal.pone.0165252
- Thomsen PF & Willerslev E (2015) Environmental DNA – An Emerging Tool in Conservation for Monitoring Past and Present Biodiversity. Biological Conservation 183: 4–18. https://doi.org/10.1016/j.biocon.2014.11.019

- Tyson, GW & Hugenholtz, P (2005). Environmental shotgun sequencing. Encyclopedia of genetics, genomics, proteomics, and bioinformatics. Edited by Lynn B. Jorde. West Sussex, UK: John Wiley & Sons.1386-1391. https://doi.org/10.1002/047001153X.g205313
- Valentini A, Taberlet P, Miaud C, Civade R, Herder J, Thomsen PF, Bellemain E et al. (2016) Next-Generation Monitoring of Aquatic Biodiversity Using Environmental DNA Metabarcoding. Molecular Ecology 25(4): 929–42. https://doi.org/10.1111/mec.13428
- Wacker S, Fossøy F, Larsen BM, Brandsegg H, Sivertsgård R, & Karlsson S (2019). Downstream transport and seasonal variation in freshwater pearl mussel (Margaritifera margaritifera) eDNA concentration. Environmental DNA, 1(1), 64-73. https://doi.org/10.1002/edn3.10
- Wilkinson M, Dumontier M, Aalbersberg I et al. (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18
- Wittwer C, Stoll S, Strand D, Vrålstad T, Nowak C, & Thines M (2018). eDNA-based crayfish plague monitoring is superior to conventional trap-based assessments in year-round detection probability. Hydrobiologia, 807(1), 87-97. https://doi.org/10.1007/s10750-017-3408-8
- Yates MC, Fraser DJ & Derry AM (2019) Meta-analysis Supports Further Refinement of eDNA for Monitoring Aquatic Species-specific Abundance in Nature. Environmental DNA. https://doi.org/10.1002/edn3.7.
- Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G & Vaughan R (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature Biotechnology 29(5): 415. https://doi.org/10.1038/nbt.1823