Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life


1 INRAE, USMB – UMR Carriel, 75bis av. de Correntz – CS 30511, FR-74203, Thonon les Bains Cedex, France
2 King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Biological and Environmental Sciences and Engineering (BESE), Thuwal 23955-6900, Saudi Arabia
3 AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Herrera Kaia, Portualdea s/n, 20110, Pasaia, Spain
4 INRAE, LR EABX, 50 avenue de Verdun Gazeinet 33612, Cestas Cedex, France
5 Slovak Academy of Sciences, Plant Science and Biodiversity Centre, Zoology Lab, Dábravska cesta 9, 84523, Bratislava, Slovakia
6 Centre of Molecular and Environmental Biology (CBMA), University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal
7 Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Campus de Gualtar, 4710-057, Braga, Portugal
8 Swiss Centre for Applied Ecostoecology (Ecosto Centre), EPFL ENAC IIE-GE, 1015, Lausanne, Switzerland
9 Institute of Marine and Environmental Sciences, University of Szczecin, Mickiewicza 16a, 70-383, Szczecin, Poland
10 Institut Pasteur, Collection des Cyanobactéries, 28 rue du Dr Roux, 75724 Paris Cedex 15, France
11 University of Geneva, Department of Genetic and Evolution, Quai Ernest Ansermet, 30 1205, Geneva, Switzerland
12 Botanischer Garten und Botanisches Museum, Freie Universität Berlin, Königin-Luise-Str. 6-8, 14195, Berlin, Germany
13 Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, PO Box 7050, SE-750 07, Uppsala, Sweden
14 University of Duisburg-Essen, Faculty of Biology, Aquatic Ecosystem Research, Univesitatsstrasse 5, 45141, Essen, Germany
15 Meise Botanic Garden, Nieuwelaan 38, 1860, Meise, Belgium
16 IRITA-Institute for Food and Agricultural Research and Technology, Marine and Continental Waters Programme. Cira de Poble Nou Km 5.5, E43540, Sant Carles de la Rápita, Catalonia, Spain
17 Royal Botanic Garden Edinburgh, Inverleith Row, Edinburgh EH3 5LR, UK
18 INRAE, U3E, 65 rue de Saint Brieuc, 35042, Rennes Cedex, France
19 University of Nantes, Laboratoire Mer Molecules Sante – EA 21 60, Faculty of Sciences, 44322, Nantes CEDEX, France
20 Institute of Oceanology, Polish Academy of Sciences, 81-712, Sopot, Poland
21 ID-Genèceodiagnostic, Campus Biotech Innovation Park, 1202, Geneva, Switzerland
22 Institut de Systémétique, Évolution, Biodiversité (ISYEB), Muséum National d’Histoire naturelle, Sorbonne Université, CNRS, EPHE, Université des Antilles, 57 Rue Cuvier CP50, 75003, Paris, France
23 University of Angers, LPG-BIAF, UMR CNRS 6112, 2 bd Lavoisier, 49045, Angers Cedex, France
24 SCIMABIO Interface, 5 rue des Quatre Vents 74200, Thonon-les-Bains, France
25 Musée national d’histoire naturelle de Luxembourg, 25 Rue Münster, 2160, Luxembourg, Luxembourg
26 Norwegian University of Science and Technology, NTNU University Museum, Department of Natural History, NO-7491, Trondheim, Norway
27 King Abdulaziz University, Faculty of Marine Sciences, Jeddah, Saudi Arabia

Corresponding author: Frédéric Rimet (frederic.rimet@inrae.fr)
Abstract

DNA barcoding and metabarcoding is increasingly used to effectively and precisely assess and monitor biodiversity in aquatic ecosystems. As these methods rely on data availability and quality of barcode reference libraries, it is important to develop and follow best practices to ensure optimal quality and traceability of the metadata associated with the reference barcodes used for identification. Sufficient metadata, as well as vouchers, corresponding to each reference barcode must be available to ensure reliable barcode library curation and, thereby, provide trustworthy baselines for downstream molecular species identification. This document (1) specifies the data and metadata required to ensure the relevance, the accessibility and traceability of DNA barcodes and (2) specifies the recommendations for DNA harvesting and for the storage of both voucher specimens/samples and barcode data.

Key Words

aquatic organisms, barcode, DNA, metadata, quality, reference library, traceability

1. Introduction

Human well-being is intimately linked to freshwater and marine ecosystems (Dudgeon et al. 2006; Wom et al. 2006; Borja et al. 2020). Although these systems support critical services to humans, such as water supply, fisheries or tourism, they increasingly face stressor impacts (Carvalho et al. 2019). Natural and especially anthropogenic alterations are introducing important pressures on rivers, lakes, coasts and seas (European Environmental Agency 2019). Beyond the services provided by these ecosystems and the need to protect their sustainability for future generations, there is an ethical need for societies to commit themselves for nature conservation.

Only 10% of European rivers have very low concentrations of micropollutants (Loos et al. 2009), which challenges citizens on the supply of aquatic ecosystem services (Boulton et al. 2016). Another alteration is related to the increasing need of human society for energy: there is a global boom of new hydropower dams in the last decades (Zarfl et al. 2015, 2019). Consequently, at global scale, only a few rivers remained free-flowing (Liermann et al. 2012; Grill et al. 2019). This increased habitat fragmentation results in a significant biodiversity loss in rivers which, in turn, impacts on ecosystem services (e.g. Arrington et al. 2010), such as food provisioning and cultural services. In marine ecosystems, new wind and wave farms, offshore drilling platforms and fish farms have been constructed in the last decades to meet energy needs. This has important impacts on biodiversity (Shield et al. 2011; Pawlowski et al. 2014c); these human activities destroy marine habitats (e.g. damage to sea floors due to trawling, coastal urban expansion, dredging and destruction of coral reefs and mangroves) or alter them (e.g. the construction of wind and wave farms, Shields et al. 2011). Human impact is particularly negative for coral reefs, mangroves and many other coastal regions (Halpern et al. 2008). However, millions of people depend upon services and functions of such ecosystems, including food, tourism and storm protection (Barbier et al. 2011). Overexploitation of marine and freshwater biodiversity, including overfishing and destruction of some habitats such as mangroves (e.g. Myers and Worm 2003) is particularly worrying. The spread of invasive species in freshwater and marine habitats has caused dramatic changes in ecosystems and biodiversity (Dextrase and Mandrak 2006; Molnar et al. 2008). Finally, global changes also impact aquatic ecosystems. In particular global warming, in combination with other pressures, such as nutrient loading, causes other detrimental effects, such as harmful algal blooming in lakes and seas (Jacquet et al. 2005; Jeppesen et al. 2005; Wells et al. 2015) or global water scarcity (Schewe et al. 2014). More specifically in marine ecosystems, global warming and acidification negatively impact marine diversity (Kroeker et al. 2013). These concerns about the state of rivers and lakes led scientists to warn about a global threat to human water security and biodiversity since more than 80% of the world’s population is exposed to such risk (Vörösmarty et al. 2010; Jenny et al. 2020).

Aquatic life and its biodiversity are crucial to ensure quality, quantity and delivery of aquatic ecosystem services (Cardinale et al. 2012; Stevenson 2014; Barbier 2017; Hammerschlag et al. 2019). For several decades, governments and transnational organisations have understood the necessity of monitoring aquatic life to set up political decisions that will help in improving or preserving aquatic ecosystems (Kopf et al. 2015). For instance, in the European legislation, several directives such as the Water Framework Directive (European Commission 2000), the European Marine Strategy Framework Directive (European Commission 2008), the European regulation on invasive alien species (European Commission 2014), the Habitats Directive (The Council of the European Communities 1992) and the European Biodiversity Strategy (European Commission 2020) were set up to protect the integrity of particular species, habitats and aquatic ecosystems. These directives are applied in each member states’ legislation. In the USA, the Clean Water Act (Copeland 2016) also sets an ambitious framework for water ecosystems preservation and, in the marine realm, many countries have implemented legislation to protect them (Borja et al. 2008).
The implementation of such directives is generally based on methodologies that assess the presence and abundance of species in freshwater or marine ecosystems (Hering et al. 2006; Borja et al. 2010; Birk et al. 2012). Most of the classical methodologies used to detect and identify species rely on morphological differences. Visual observations or optical devices (binocular magnifiers or microscopes) enable experts to determine taxon and establish fauna or flora identification lists. However, such methods are time-consuming, require a high level of expertise to reliably identify the organisms and are, in some cases, unable to distinguish amongst closely-related or morphologically-indistinguishable species (i.e. cryptic species). Identification of immature stages (e.g. different ontogenetic stages, such as eggs, juveniles, planktonic larvae) is also often impossible, especially for small organisms. The use of molecular, DNA-based tools to identify species overcomes this problem as long as DNA is preserved in the sample. DNA barcoding, the use of short, standardised gene sequences for species identification (Hebert et al. 2003) has become the most versatile and universally-applicable method since the adoption of standard DNA markers was agreed upon within the scientific community (e.g. COI in animals (Hebert et al. 2003), ITS in fungi (Schoch et al. 2012), MatK, rbcL (CBOL Plant Working Group 2009) or ITS (Li et al. 2011) in plants. The development of open access analytical tools and comprehensive data repositories (e.g. BOLD, the Barcode of Life Data Systems and Barcode Index Numbers (BINs)) has fast-forwarded the advancement of biodiversity assessments using DNA barcoding (Ratnasingham and Hebert 2007; Ratnasingham and Hebert 2013). The method relies upon DNA reference libraries that link species-specific taxonomic classification to a reliable reference sequence (the barcode). With the entry of new sequencing techniques, the concept has been extended to identify specimens present in bulk or environmental samples (rather than single individuals) and dubbed DNA metabarcoding (Pompanon et al. 2011). Both DNA barcoding and metabarcoding have been successfully applied to key aquatic organisms used for ecosystem assessment: such as diatoms (Kermarrec et al. 2013; Zimmermann et al. 2014b), Foraminifera (Pawlowski and Lecroq 2010; Pawlowski et al. 2014a, b, 2016; Frontalini et al. 2020), ciliates, macroinvertebrates and particularly aquatic insects (Hajibabaei et al. 2012; Elbrecht and Leese 2015), marine benthic fauna (Aylagas et al. 2014; Lobo et al. 2017), fish (Hänfling et al. 2016; Pont et al. 2018), aquatic oligochoaetes (Vivien et al. 2016, 2019), macroalgae and aquatic angiosperms (Scriver 2015; Akita et al. 2019, 2020). The international initiative iBOL (International Barcode of Life), currently represented by 35 member nations, is a research alliance committed to building and extending existing DNA barcode reference libraries, informatics platforms, analytical protocols and pipelines for assessments of biodiversity using molecular tools (https://ibol.org/).

An Achilles heel of DNA barcoding and metabarcoding is the taxonomic coverage and the data quality of the barcode reference libraries. Weigand et al. (2019) give an overview of gaps in the barcode libraries for species used in aquatic biomonitoring in Europe and show that, while more than 80% of all fish species are barcoded, only 26% of the marine invertebrates are covered. They also argue for the need for effective quality assurance and quality control of barcode reference libraries. With this in mind, it is important to develop and follow best practices to ensure that the quality and traceability of the metadata associated with the reference barcodes used for identification is optimised. As early as 2005, data standards for reference DNA barcode records to be deposited in INSDC (International Nucleotide Sequence Database Collaboration) were given (Hanner 2005 revised in Hanner 2009). However, these standards were customised for INSDC, had a generalist nature and missed the specificity required for different organisms and ecosystems. Some guidelines for specific organisms were made for diatoms (e.g. Zimmermann et al. 2014a) and were standardised at European level (European Standardisation Committee 2018) and then translated into each member state’s language (e.g. in French: Afnor 2018). Similarly, some specific standards for data and protocols for fish were proposed in the framework of the Fish Barcode of Life campaign (Ward et al. 2009; Ward 2012). We focus on quality standards before the molecular generation of DNA barcodes commences, i.e. pre-PCR and sequencing. The objectives of this document are: (1) to specify the data and metadata required to ensure the relevance, the accessibility and traceability of DNA barcodes and (2) to specify the recommendations for DNA harvesting and for the storage of both voucher specimens/samples and barcode data.

2. Procedures

This section details the procedures for the storage of vouchers, DNA material, DNA-barcodes and for the harvesting of DNA.

2.1 Storage of voucher specimens

The first step in establishing a reference barcode is the correct morphological identification of specimens from which DNA will be isolated. The physical vouchers of the identified biological material should be deposited in a recognised and accessible natural history collection and be accompanied by a unique collection number (Vollmar et al. 2010; Blagoderov et al. 2012). Duplicates of vouchers could be deposited in an alternative recognised collection(s) to reduce the risk of losing vouchers and increase accessibility. Labels containing obligatory metadata must be attached to all parts of the specimens or preparations with the specimens included (such as permanent microscopic slides). Depending on the organism group, the physical vouchers can have different forms and the following sections (2.1.1
to 2.1.10) give for each organism, the obligatory and the recommended vouchers for the different types of biological materials usually used for DNA harvesting.

In the following sections, depending on the organisms considered, we recommend that vouchers are stored frozen. Various temperatures are presently used (-20 °C, -40 °C, -50 °C, -80 °C), but there is a lack of long-term comparisons (e.g. 100 years) of the impact of freezing temperatures on DNA conservation. Thus, we recommend that frozen vouchers should be stored at least -20 °C (or temperatures from -20 °C to -80 °C).

2.1.1 Cyanobacteria

Different kinds of biological material can be used:
- Dry specimens:
  The obligatory voucher must be the thallus or large colonial structure built by cyanobacteria. The recommended voucher can be dried treated samples.
- Monoclonal isolate in culture:
  The obligatory voucher must be a living culture with a unique strain number kept in a culture collection with appropriate light source (quality and quantity of daylight).
- Monoclonal and axenic isolate in culture:
  The obligatory voucher must be a culture deposited in two official collections (e.g. PCC [Pasteur Culture Collection] in Paris, France, ATCC [American Type Culture Collection] in the USA, NIES-MCC [Microbial Culture Collection at the National Institute for Environmental Studies] in Tsukuba, Japan); those collections will give a unique number that is requested for publishing in official journals.

2.1.2 Diatoms

Different kinds of biological material can be used for diatoms:
- Culture:
  The obligatory voucher must be biomass of the monoclonal culture kept at -80 °C or fixed in ethanol or fixed in a buffer and a labelled permanent microscope preparation of cleaned culture (frustules and valves). The recommended voucher can be a living culture with a unique strain number (kept in a culture collection), dried treated material, scanning electron microscope (SEM) stubs, loose dried unoxidised material, permanent slide. The recommended documentation are photographs of the alive cells and valves from slides and SEM stubs.
- Single cell:
  The obligatory voucher must be light and/or electron microscope photographs showing diagnostic details of the cell. The recommended voucher can be a permanent microscopic slide with the frustule, if it has not been destroyed after extraction.
- Environmental sample:
  The obligatory voucher must be raw material kept at -80 °C or fixed with ethanol (> 70%) or formaldehyde or buffer and a labelled permanent microscope preparation of the cleaned culture (frustule and valves). The recommended voucher can be dried treated material, scanning microscope stubs, loose dried unoxidised material, photographs.

2.1.3 Other microalgae

We propose one kind of biological material that can be used for other microalgae (other than diatoms and cyanobacteria); however, other options may be considered:
- Culture:
  The obligatory voucher must be a monoclonal culture kept at -80 °C or fixed in ethanol or fixed in a buffer. The recommended voucher can be a living culture with a unique strain number (kept in a culture collection), dried treated material, light and/or electron microscope photos showing diagnostic details of the cell. Some microalgal taxa can also be cryopreserved (e.g. Stock et al. 2018).

2.1.4 Macroalgae and aquatic angiosperms

One kind of biological material can be used for macroalgae and aquatic angiosperms:
- Specimen kept dry or wet:
  The obligatory voucher must be a voucher dried on a herbarium sheet. The recommended vouchers can be parts of the plant preserved wet (ethanol or formaldehyde) for anatomical or detailed morphological observations and/or a living culture with a unique strain number (kept in a culture collection).

2.1.5 Foraminifera

Different kinds of biological material can be used for Foraminifera:
- Single cell with mineral test:
  The obligatory voucher must be electron microscope photographs showing details of the test. Since tests are destroyed during extraction, paratypes should be kept on micropaleontological slides. The recommended voucher can be dried tests of paratypes stored in micropaleontological slides at room temperature.
- Single cell with organic wall or naked:
The obligatory voucher must be light microscope photographs showing important details of the cell. Since cells are destroyed during DNA extraction, paratypes should be fixed in formalin and, for long term storage, be transferred into 70% ethanol. The recommended voucher can be tested fixed in 4% formalin and permanently stored in 70% ethanol.

- Environmental sample: 
The obligatory voucher must not be treated material of the environmental sample stored at -20 °C or -80 °C. The recommended voucher can be single cells isolated from the environmental sample, kept on slides or fixed and stored in 70% ethanol.

2.1.6 Macrinovertebrates, including crustaceans, echinoderms, insects, sipunculids and cnidarians

Different kinds of biological material can be used:
- Specimen frozen, dried or preserved in ethanol:
The obligatory voucher must be the hard exoskeleton dried for terrestrial taxa (e.g. pinned insects) or a specimen kept in ethanol (70–96%) or frozen (-20 °C). Slide mounts of part or all of the whole specimen in permanent medium is recommended, if needed for morphological observations. The recommended voucher can also be tissue samples of specimens kept at -80 °C or -20 °C if the obligatory voucher is preserved in 70–96% ethanol.

2.1.7 Annelids

One kind of biological material can be used (Timm and Martin 2015; Vivien et al. 2017):
- Specimen preserved in absolute ethanol at -20 °C:
The obligatory voucher must be the anterior parts of specimen (at least the first 15 segments) kept in ethanol (> 80%) or in 4% formalin or mounted on a slide in a permanent medium. The recommended voucher used for subsequent genetic analyses can be a tissue sample of the obligatory voucher kept in absolute ethanol at -20 °C or -80 °C.

2.1.8 Molluscs

Two kinds of biological material can be used:
- Specimen preserved dry:
  For shellless specimens, the obligatory voucher must be a shellless specimen preserved dry. The recommended voucher can be a separate tissue sample, high-resolution imaging data for shell surface and internal organisation (e.g. scanning electron microscopy, microtomography) and 3D model. For shell-less specimens (e.g. slugs), no ideal method for dry vouchers exists, but comprehensive imaging should be performed on living specimens prior to storing specimens in a wet collection.
- Specimen preserved wet:
The obligatory voucher must be a specimen stored wet in a preservative suitable for morphology and genetics (e.g. 80% ethanol, propylene glycol) or kept at -20 °C or -80 °C. The recommended voucher can be specimen tissues stored in a preservative suitable for morphology and genetics, comprehensive imaging data, permanent microscopic slide(s) with genital apparatus, mouth parts or other diagnostic morpho-taxonomic characters. Due to their high water content, wet preserved molluscs dilute the preservation liquid. As such, the initial preservative should be renewed after 24 h and a generally fair preservative to tissue ratio respected (e.g. > 5:1).

2.1.9 Fish

Two kinds of biological material can be used:
- Entire specimen or part of body preserved in ethanol or frozen:
  Obligatory voucher must be entire body or part of body preserved in ethanol (> 80%) or frozen (-20 °C).
- Tissues samples preserved in ethanol or frozen or scales preserved dry:
The recommended voucher can be tissue samples of specimens kept at -80 °C or in ethanol (> 70%) or scales preserved dry.

2.1.10 Aquatic fungi

Two kinds of biological material can be used:
- Population from environmental sample:
The obligatory voucher must be a sample collected on a polycarbonate filter (0.6 μm pore-size) and stored at -20 °C or -80 °C. The recommended voucher can be a photograph.
- Obligate parasites:
The obligatory voucher must be a living culture isolate with its host. The recommended voucher can be raw cultures material kept frozen (-20 °C or 80 °C), photographs.

2.2 DNA harvesting and banking

DNA can be extracted from a variety of biological materials depending on the organism group (cultures, tissues, populations, natural samples of mixed organisms etc.). As Table 1 illustrates, the preferred method to extract DNA differs amongst organism groups and, for some organisms, multiple methods are possible. In common for most groups is the downstream use of PCR and a choice of Sanger or High Throughput Sequencing (HTS) to generate the DNA barcode depending on the state of the specimen (e.g. age, mixed sample etc.). However, also shallow shotgun sequencing (i.e. genome skimming) is used to generate reference barcodes (Alsos et al. 2020).

It is of great importance to store an aliquot of extracted DNA permanently to make DNA available for future research. The aliquot should be stored permanently at -20 °C or -80 °C in an established DNA bank or a biological specimen repository, with links to the corresponding metadata.

https://mbmg.pensoft.net
### Table 1. Biological material from which DNA can be harvested and recommendations for procedures to isolate DNA to generate DNA barcodes for aquatic organisms.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Biological material</th>
<th>Recommended method for DNA isolation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cyanobacteria</td>
<td>Dried specimen in herbarium</td>
<td>DNA extraction can be performed with commercial kits optimised for plants.</td>
<td>Wright et al. (2001)</td>
</tr>
<tr>
<td></td>
<td>Isolated specimen from freshwater and brackish water (filaments for example)</td>
<td>A base phenol/chloroform extraction is feasible, but even boiling/freezing to release the DNA or boiling individual filaments in the reaction mix tube in the PCR machine prior to adding the DNA polymerase will work.</td>
<td>Laamanen et al. (2001, 2002)</td>
</tr>
<tr>
<td></td>
<td>Culture of monoclonal isolate, axenic or not</td>
<td>Several DNA extraction methods are available, but the one obtained is Shih et al. (2013) which was tested successfully on a wide diversity of cyanobacteria.</td>
<td>Shih et al. (2013)</td>
</tr>
<tr>
<td>Diatoms</td>
<td>Monoclonal culture from single algal cell isolation</td>
<td>Cell isolation from an environmental sample, which is then grown in culture. DNA extraction from the culture, PCR and Sanger sequencing. Culture does not need to be axenic, but must host a single diatom taxon. Adopt this methodology for cultivable species.</td>
<td>Abarca et al. (2014); Zimmermann et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>Single cell from fresh environmental sample*</td>
<td>Individual cell isolation from an environmental sample. DNA extraction, PCR and Sanger sequencing from the cell. This methodology is mostly suitable for large-celled species.</td>
<td>Hamilton et al. (2015); Frankovich et al. (2018); Skibbe et al. (2018)</td>
</tr>
<tr>
<td></td>
<td>Population from fresh or ethanol (&gt; 70%) preserved environmental sample</td>
<td>DNA extraction of environmental sample, PCR, cloning and Sanger sequencing. Adopt this methodology for species in high abundance in an environmental sample. Alternatively, high-throughput sequencing of the environmental sample, with subsequent bioinformatics and phylogenetic analysis to isolate the target species barcode. Adopt this method when target species are relatively well represented in an environmental sample, but are difficult or impossible to cultivate.</td>
<td>Khan-Hureua et al. (2016); Rimet et al. (2018)</td>
</tr>
<tr>
<td>Other microalgae</td>
<td>Culture of monoclonal isolate, pure or not</td>
<td>DNA extraction from the culture, PCR and Sanger sequencing. Culture does not need to be axenic, but must host a single algal taxon. Adopt this methodology for cultivable species.</td>
<td></td>
</tr>
<tr>
<td>Macroalgae</td>
<td>Specimen silica-dried, from herbarium specimen or preserved in ethanol (&gt; 70%)</td>
<td>DNA extraction of a clean portion of the thallus (with as little epiphytes as possible), PCR, Sanger sequencing. Different barcode markers are routinely applied for different groups of macroalgae.</td>
<td>Saunders, (2015); McDevitt and Saunders (2009); Saunder and McDevitt (2012); Hall et al. (2010); Lelaitet et al. (2014)</td>
</tr>
<tr>
<td></td>
<td>Population from environmental sample</td>
<td>DNA extraction, PCR and high-throughput sequencing of the environmental sample, with subsequent bioinformatics and phylogenetic analysis to isolate the target species barcode.</td>
<td>Akita et al. (2019, 2020)</td>
</tr>
<tr>
<td>Aquatic angiosperms</td>
<td>Specimen silica-dried, from herbarium or environmental sample</td>
<td>DNA extraction of a clean portion of the thallus (with as little epiphytes as possible), PCR, Sanger sequencing or high-throughput sequencing of the environmental sample.</td>
<td>Qubramanzadeh et al. (2013), Wang et al. (2010), Lucas et al. (2012), Sirever et al. (2015)</td>
</tr>
<tr>
<td>Foraminifera</td>
<td>Single cell isolated from environmental sample</td>
<td>Sediment samples are sieved over 500, 250, 125 and 63 μm sieves and stored separately in containers filled with seawater. Samples are inspected under a stereomicroscope and living foraminifera (distinguished by brownish, yellowish, reddish or greenish colour of cytoplasm and pseudopodial movement) are isolated. DNA extraction of single specimens in up to 500 μm of guanidine lysis buffer. For larger specimens, commercial kits adapted for plants can be used for DNA extraction. Semi-nested or nested PCR of the 18S rRNA barcode fragment for foraminifera and Sanger sequencing. Some species exhibit high intra-individual variability in their 18S rRNA genes and need to be cloned prior to sequencing.</td>
<td>Pawlowski (2000); Pawlowski and Holzmann (2014); Darling et al. (2016)</td>
</tr>
<tr>
<td></td>
<td>Population from environmental sample</td>
<td>DNA extraction using commercial kits adapted for soil. PCR with specific foraminiferal primers for 18S barcoding regions, high-throughput sequencing, bioinformatic and phylogenetic analyses to identify metabarcode.</td>
<td>Moss et al. (2016); Leucoq et al. (2011); Lejerowicz et al. (2015); Pawlowski et al. (2014b, c)</td>
</tr>
<tr>
<td></td>
<td>Culture from single or few cells from environmental sample</td>
<td>Only a few foraminiferal species can be cultured. Foraminifera have a complex life cycle alternating between sexual/asexual reproduction and only a few species are able to maintain repeated asexual reproduction under laboratory conditions. Single cells are isolated from cultures and extracted with guanidine lysis buffer. Semi-nested PCR, cloning (depending on intra-individual polymorphism), Sanger sequencing.</td>
<td>Siemensen et al. (2017); Barras et al. (2009); Apollé-Ho-Perret-Gentil et al. (2015); Holmzmann et al. (2018)</td>
</tr>
<tr>
<td>Macrourvertebrates, including arthropods, echinoderm, spumiculids and cnidarians</td>
<td>Specimen fresh or preserved dried/frozen, in ethanol (&gt; 70%) or frozen</td>
<td>Destructive DNA extraction of tissue sample, without destroying diagnostic features in the remaining voucher (e.g. leg, muscle tissue), PCR, high-throughput or Sanger sequencing.</td>
<td>Marine metazoas: Steenke et al. (2016); Insect: Porco et al. (2010)</td>
</tr>
<tr>
<td></td>
<td>Non-destructive DNA extraction of a complete specimen, PCR, high-throughput or Sanger sequencing.</td>
<td></td>
<td>Haran et al. (2018)</td>
</tr>
<tr>
<td>Annelids</td>
<td>Specimen preserved in ethanol (&gt; 80%) at ~20 °C or in neutral buffered formalin or fresh/frozen**</td>
<td>DNA extraction of entire specimens or a fragment of specimens using guanidine lysis buffer or a commercial kit, PCR and Sanger sequencing or high-throughput sequencing of genetically-tagged specimens.</td>
<td>Vivien et al. (2017, 2020)</td>
</tr>
</tbody>
</table>
Metabarcoding and Metagenomics 5: e58056

<table>
<thead>
<tr>
<th>Organism</th>
<th>Biological material</th>
<th>Recommended method for DNA isolation</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molluscs</td>
<td>Specimen / tissue preserved in ethanol (&gt; 70%) or propylene glycol</td>
<td>DNA extraction, inhibitor removal, PCR and then high-throughput sequencing or Sanger sequencing.</td>
<td>Zhang et al. (2010)</td>
</tr>
<tr>
<td></td>
<td>Specimen / tissue preserved dry</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fish</td>
<td>Specimen or tissue preserved in ethanol (&gt; 70%) or frozen, scales preserved dry</td>
<td>DNA extraction of muscle tissue sample (usually white muscle taken from under a scalpel-cut skin flap on the right side of the fish), of thin tissues or scales, PCR and high-throughput or Sanger sequencing.</td>
<td>Ward (2012), Marchand et al. (2018)</td>
</tr>
<tr>
<td>Aquatic fungi</td>
<td>Zoosporic fungi: Population from environmental sample</td>
<td>Enzymatic digestion and lysis buffer to digest the fungal chitin wall. Incubation with proteinase K and SDS for protein digestion and DNA extraction. DNA purification method using silica-membrane columns. PCR, cloning and sequencing.</td>
<td>Lefèvre et al. (2010); Jobard et al. (2012)</td>
</tr>
<tr>
<td></td>
<td>Filamentous fungi and yeast: isolate kept in culture</td>
<td>Disruption of filamentous fungal cell walls with glass bead method, followed by digestion using proteinase K. Rinsing fungal mycelia or yeast cells with pure water to remove potential PCR inhibitors, followed by thermolysis at 85 °C in lysis buffer, DNA amplification and extraction.</td>
<td>Gontia-Mishra (2014); Zhang et al. (2010)</td>
</tr>
</tbody>
</table>

*Methodology is very difficult as it is challenging to get pictures and PCR products of the same cells.*

** Fixation of aquatic oligochaetes and most soft-bodied organisms in a low concentration of ethanol can lead to a fragmentation/disintegration of specimens (Vivien et al. 2018, 2020b). For in situ fixation of large (> 2-3L) sediment samples, it is recommended to use neutral buffered formalin (final concentration of 4% neutral formaldehyde) instead of ethanol to avoid the destruction of oligochaete specimens (Vivien et al. 2018, 2020b). Seeding should be performed within 4 weeks after sampling and material retained should be stored in absolute ethanol at -20 °C (Vivien et al. 2018).

(see Section 3 below). To secure its visibility and accessibility, we recommend that the repository used is affiliated with a national or international network such as RARE, the French Agronomic Biological Resources Center network (https://www.agrobrc-rare.org/agrobrc-rare_eng/; Mougin et al. 2018), the Global Genome Biodiversity Network -GGBN- (http://www.ggbn.org, Droge et al. 2014), the European Research Infrastructure DiSSCO (Distributed System of Scientific Collections, https://www.dissco.eu/) or the Consortium of European Taxonomic Facilities (CETA, https://cetaf.org) which link institutions hosting collections that preserve genomic resources. Biological vouchers (reference specimens) should be stored in a publicly-accessible scientific collection that practises regular loans of material for scientific use (e.g. museums and natural history collections, national institutes collections and botanical gardens) (see details in Section 2.3).

### 2.3 Storage of DNA barcodes

We strongly recommend DNA barcodes and associated metadata to be stored digitally in public, open-access databases. Examples include general databases like BOLD (Ratnasingham and Hebert 2007) which links metadata and images directly to the barcode, but also ENA (Amid et al. 2019) or NCBI GenBank (Benson et al. 2013) can do so with the appropriate tags. DNA barcodes can be also stored in more specialised databases that are frequently curated by experts such as SILVA (Quast et al. 2013), PR2 (Guilhou et al. 2013), EukRef for eukaryotes (del Campo et al. 2018), now integrated in PR2 (https://pr2-database.org/eukref/about/), UniEuk (https://unieuk.org) PhytoRef for algae (Decelle et al. 2015), Algaterra for microalgae (Kusber et al. 2012, 2020+), Diat.barcode for diatoms (Rimet et al. 2019), UNITE for fungi (Nilsson et al. 2019), Greengenes (DeSantis et al. 2006) and RDP (Cole et al. 2014) for bacteria, archaea and fungi. However, while most of these specialised databases practise taxonomic curation of the DNA sequence data, some are less concerned about keeping all the necessary metadata to ensure optimal traceability and, thus, reliability of the reference database.

### 3. Associated metadata

We consider the DNA sequence as primary data and all accompanying information as metadata. (See supplementary file for tabular overview of the below listed metadata).

#### 3.1 General remarks about metadata

A reference barcode must be accompanied by a minimum set of metadata. The metadata, including photographs, must be stored in digital and open-access databases, such as those listed in Section 2.2 or in a publicly accessible collection database of the storing institution.

It is important that the metadata is stored in non-proprietary formats (e.g. text documents in .txt, images in .tif) and that such databases comply with FAIR Data principles (Findable, Accessible, Interoperable, Re-usable, see Wilkinson et al. 2016) and the Biodiversity Information Standards, such as the Darwin Core, Audubon Core, ABCD 2005 standard (https://www.tdwg.org).

The metadata must be linked to the barcode via a unique identifier assigned by the database where the sequence is stored (e.g. accession number in ENA or in GenBank). If voucher material or culture strains are available in natural history or institutional culture collections, these should be linked to the sequence.

As detailed in the following sections, metadata should include information on the DNA marker, the strain cultivation, the natural sample, the taxonomy name, the identification, the sampling location, the voucher location and the barcode authors.
3.2 Categories of metadata

3.2.1 Biological material metadata

The metadata listed below give the obligatory and recommended items that ensure the traceability of the biological material used for DNA harvesting (see Section 2.1).

### Biological specimens and environmental samples

#### Obligatory metadata

1. Location of the sampling site
   a) Geographical coordinates: for example, expressed in decimal values in WGS84 or in a different, specified geographical positioning system.
   b) Country according to the ISO 3166 standard, accepted name of ocean or sea.
   c) Name of the locality.

#### Remarks:

- For species of heritage interest, Red List species or endangered species, national or regional regulations might ask not to reveal precise location coordinates in order to protect their populations. These regulations should be followed and the exact locality information hidden.
- In some cases, exact coordinates are not available (e.g. when older museum specimens are used; here a georeference of the locality plus an estimated uncertainty in metres can be added).

2. Date of sampling, preferably in ISO-format (YYYY-MM-DD).
3. Name of person who collected the specimen.
4. Photo(s) of the voucher specimen showing diagnostic features, including scale(s).
   a) Macroscopic organisms: whole specimen or specified parts important for morphological identification. For fish, photos should be taken of the left side of the specimen. For specimens in which the morphology can be altered during storage (e.g. dry storage of molluscs) or preservation method (e.g. ethanol for anthozoans, molluscs, polychaetes and sipunculids, for example, Manuel 1988; Saiz Salinas 1993; Nygren et al. 2011), high resolution photos should be taken before preservation. Relaxing chemicals may be used to facilitate imaging of live specimens to assure visibility of key morphological parts (e.g. proboscis eversion in polychaetes; Nygren et al. 2011; Bonyadi-Naeini et al. 2016).
   b) Microscopic organisms: light microscopy and/or electron microscopy micrographs must show all the important details of the cells or the colonies that are necessary for identification.
5. Preservation status and fixative used to preserve the sample.
6. Conditions for access to the material: legal requirements for use or re-use of DNA/material resources with eventual references to national/international laws. A material transfer agreement might be necessary to regulate exchange.
7. Reference of the document (published article, report etc) linked to the deposition of the barcode sequence.

#### Recommended metadata

1. Environment (ecosystem) at sampling site (e.g. Lake, river, swamp, tidal flat, open sea, groundwater, hyporheos, mangrove, lagoon, estuary, deep sea, rocky shore, coral reef, etc.).
2. Substrate (rock, macrophyte, sediment, hot vent, interstitial, etc.).
3. Habitat (e.g. plankton, epipelon, epilithon, epipsammon, tychoplankton, alluvial region, porous or karstic aquifer, sea floor, pelagic, benthic, intertidal, subtidal, etc.).
4. Sampling elevation (m a.s.l.).
5. Sampling depth (m).
6. Sampling device or sampling protocol.
7. Photos of the sampling site.
8. Environmental measurements: luminosity, pH, conductivity, salinity, temperature, sediment’s grain size, organic matter content and redox potential.
9. Main ecological function(s) of the specimen (if known). For instance, already existing ecological classifications, such as FAPROTAX (Louca et al. 2016) and Tax4Fun (Abhauer et al. 2015) for bacteria, the classification of Reynolds et al. (2002) and Padisák et al. (2009) for phytoplankton and Rimet and Bouchez (2012) for diatoms, for macro-invertebrates the classification of Ússéglio-Polatera et al. (2000) or for plants, the one of Kattge et al. (2011).
10. Photos should carry the name of the photographer and associated licence, preferably a Creative Commons Licence that allows usage by third parties.
11. The FAO fishery area where the sampling was done (for marine taxa).

#### Cultures

##### Obligatory

1. Culture medium (recipe of medium used for cultivation).
2. Name of person who isolated the starting cell.
3. Date of isolation (date the uni-algal culture was established by isolating a cell from the environmental sample).
4. Date of harvesting (date the culture was harvested to extract DNA).
5. Photo(s) showing diagnostic features of the organism.

##### Recommended

1. Culture condition (light intensity, light cycles, temperature, humidity etc.).
2. Strain identifier (name or number that uniquely identifies the cultured strain in the collection).

#### 3.2.2 Taxonomic information

To ensure universal practices across countries, taxonomic information must follow the international no-
menclatural rules. Therefore, we recommend following the European standard (CEN 2014) for taxonomic identification. This standard includes, in particular, the following items:

**Obligatory data**
1. The most reliable identification to the lowest possible taxonomic rank.

**Recommended data**
1. Genus names should include the name of the author(s) with year of publication of its original valid (algebra, bacteria, fungi and plants) or available (animals) publication. Genus names should be written in italics.
2. Species epithets should include the name of the author(s) with year of publication of its description or combination (see above). Species epithets should be written in italics, while “sp.” should not.
3. If applicable, include name of infraspecific taxon (subspecies, variety, forma etc.) and citation of the author(s) with year of publication of its description or combination (see above).
4. Further notes on taxon status (e.g. phylogenetic affiliation, statements on the taxon concept adopted (i.e. whether a narrow/strict (sensu stricto) or a broader (sensu lato)). If necessary, a taxon may be given including link to a published circumscription (Berendsohn 1995).

3.2.3 Identification metadata

**Obligatory data**
1. Name of the person who identified the specimen.
2. Date on which identification was made.
3. Identification method (e.g. morphology, BOLD ID engine, NCBI BLAST, etc.).
4. If applicable, use accepted terms to indicate uncertainty in the identification (e.g. aff., cf., sp., etc.).

**Recommended data**
1. Name of the person who verified the identification.
2. Identification history including dates for identifications.
3. Reference to literature used to identify the material, specifying figures and pages that helped the identification.

3.2.4 Vouchers metadata

**Obligatory**
1. Full name and acronym of the institution, natural history collection or of culture collection where the voucher is deposited (preferably according to Index Herbariorum for algae, fungi and plants).
2. Unique identifier given to the voucher specimen by the collection where it is deposited and permanently physically associated with the specimen via indelible labelling.
3. Voucher condition: whole specimen, body part, frozen material, living culture etc.
4. Voucher status: regular specimen, holotype, paratype, syntype, lectotype, paralectotype, neotype etc.

**Recommended**
1. Full name and acronym of the institution, natural history collection or culture collection where a duplicate voucher is deposited.
2. Unique identifier given to the duplicate voucher specimen by the collection where it is deposited.
3. Institution or collection where the original DNA sample is deposited.
4. Unique identifier given to the DNA sample and given by the collection where it is deposited.
5. Long-term stable identifiers linking directly to institutions in a publicly-available database (see Güntsch et al. 2017).

3.2.5 DNA marker metadata

**Obligatory**
1. Name and abbreviation of DNA marker (e.g. 18S, 16S, rbcL, cox1 …) following a standardised nomenclature (e.g. nomenclature given in: NCBI, VGNC, HGCN).
2. Name of barcode that includes details of the region inside a particular marker (e.g. 18S v4, COI-5′ etc.).
3. Forward and reverse primer with name and sequence from 5′→ 3′ direction (cite first publication, if applicable). If more of the DNA marker than the barcode region is sequenced (i.e. a longer sequence), all primers used should be listed.
4. Sequencing technology must be given, with the brand and name of the sequencer (e.g. Sanger se-
quencing on an Applied Biosystems 3730xl DNA Analyzer; high-throughput-sequencing on an Illumina MiSeq v2).
5. Sequence quality information (e.g. Phred score, Pherogram, fastq file, coverage or similar means to assess the quality of the stored sequences or sequencing runs).
6. Barcode generators: Person(s) who actively participated in generating the barcode.

Recommended
1. DNA extraction method including protocol, brand name of the kit, if applicable.
2. PCR Mix: recipe, reaction volume, name of brand mix, if applicable.
3. PCR Protocol: protocol used for PCR amplification including duration of each step (denaturation, annealing, elongation), annealing temperature and number of cycles. The name and brand of the cycler.
4. Company or laboratory that performed the sequencing.
5. For high-throughput sequencing: software and protocol used for sequence assembly.

4. Conclusions

Weigand et al. (2019) highlighted that particular care must be taken regarding the quality assurance/quality control of the reference barcode records to be produced, as failure to do so will limit their application, render them useless or even introduce wrong outcomes. The procedure and description of metadata associated with reference sequences given in this document are steps in this direction. We do not give a procedure to control the quality of the barcode, but we hope our descriptive overview and recommendations will enable standardisation and best-practice in the production and curation of reference DNA barcodes. The metadata provide key information for users of barcode reference libraries which could be time-saving in data-analysis processes. Some of them, such as ecological functions or ecological classifications, can lead to more direct utilisation of metabarcoding in functional ecological studies.

This document might appear to set ambitious targets for those producing barcodes; however, we believe that this minimum is necessary to ensure quality in barcode reference libraries and thus provide trustworthy results for DNA barcoding and metabarcoding. Moreover, a general focus of sustainable development and energy-saving measures should be considered by the collections hosting the vouchers.

Finally, as an overall philosophy, we wish to encourage forward thinking on the format and the contents of barcode libraries and on the need for a secure access to the invaluable genetic information therein, including the information linked to specimens from which the DNA originated.

5. Glossary

Arthropods: Multicellular animals of the phylum Arthropoda, including insects, arachnids, myriapods and crustaceans.
Barcode: See DNA barcode.
Base pair: Pair of complementary cross-linked nucleotides that are the building blocks of the DNA double helix.
Biological specimen: An organism of any kingdom (animal, plant or fungi) or part of an organism. Can be living (‘living specimen’), frozen, dried (e.g. herbarium material, pinned insects, fish scales) or preserved in liquid preservatives (e.g. entire fish in ethanol: ‘preserved specimen’).
Culture: ex-situ cell culture. As a clonal culture derived from one isolated cell from the environment
Cultivator: Person responsible for the cultivation of a strain.
Cyanobacteria: phyllum of free-living photosynthetic bacteria.
Diatoms (Bacillariophyta): Group of unicellular algae, some of which form filaments or colonies, with cell walls made of silica. They are major contributors to primary productivity worldwide and are often used in ecological assessment.
DNA barcode: A stretch of DNA from a universally-accepted DNA marker that uniquely identifies specimens to species, in the context of DNA-based identification, often called just ‘barcode’.
DNA marker: Name of the coding or non-coding region (e.g. gene, spacer region) within the genome from which the barcode has been sequenced. The naming of the coding or non-coding regions should follow standard scientific practice.
ENA: European Nucleotide Archive (www.ebi.ac.uk/ena).
Environmental sample: Collection of a portion of a natural environment (water, sediment, soil or air). It contains DNA from organisms living in this environment.
Fish: A group of vertebrates containing jawless fish (Agnatha), cartilaginous fish (Chondrichthyes) and bony fish (Osteichthyes).
Fungi: group of heterotrophic eukaryotes including zygo sporic, filamentous and yeast forms.
Foraminifera: Group of unicellular heterotrophic or mixotrophic eukaryotes with organic theca or agglutinated or mineral test (rarely naked) living in all marine environments and also found in freshwater and soil. Foraminifera are used as bioindicators and can give information on pre-anthropogenic conditions as they fossilise.
GenBank: National Institutes of Health (USA) genetic sequence database, an annotated collection of all publicly available DNA sequences.
Habitat: Specific environment in which an organism lives.
HGCN: Human Gene Nomenclature Committee.
Insects (class Insecta): Hexapod invertebrates within the phylum Arthropoda (e.g. beetles, flies, odonates), characterised by a chitinous exoskeleton, a three-part body (head, thorax and abdomen), three pairs of jointed legs, compound eyes and one pair of antennae.

Isolate: A population of cells isolated from a natural population in order to be cultured and studied. The term is usually applied in microbiology.

Isolator: person responsible of the isolation of the cell from which the clonal culture was established.

Macroalgae: Macroscopic algae, comprising red (Rhodophyta), green (Viridiplantae) and brown (Phaeophyceae) lineages and forming ecologically-important primary producers in marine (all three lineages) and freshwater (green algae, mainly) ecosystems.

Metabarcoding: An identification method that enables identification of a mixture of organisms in a sample using short DNA sequences and high-throughtput sequencing.

Molluscs: An organism group referring to the taxa Gastropoda (snails and slugs), Bivalvia (e.g. clams, scallops, mussels), Polyplacophora (chitons), Cephalopoda (e.g. squids, octopus), Scaphopoda (tusk shells), Aplacophora and Monoplacophora.


Obligate parasite: a living organism which depends on a host to complete its lifecycle.

Oligochaetes: Class of the phylum Annelida. The principal aquatic oligochaete families are Naididae (Naidinae and Tubificinae), Enchytraeidae, Lumbriculidae, Haplotaxidae and Propappidae. In addition, the Lumbriculidae family includes aquatic and amphibious species.

Pherogram: Graphical account of the results from Sanger sequencing where each nucleotide is represented by a single peak and the sequence of peaks correlates to the DNA sequence of the sample analysed.

Primer: Strand of nucleic acids that serves as starting point for DNA replication.

PCR: Polymerase Chain Reaction; process used for the amplification of a target region of DNA.

PCR primer: synthesised short single-stranded nucleic acids that serve as starting point for DNA synthesis in the PCR-reaction.

Taxon (plural taxa): Taxonomic unit, for example family, genus or species. In systematics, it designates a unit to which living beings are assigned according to certain criteria. Each known taxon has a scientific name, nomenclatural type and a circumscription.

Taxonomy: Taxonomic concept specified by the scientific name and the reference in which its name is used.

Taxonomic backbone: Index of published taxon names which is used by databases to automatically cross-reference name entries.

VGN C: Vertebrate Gene Nomenclature Committee.

Voucher: Physical specimen of a sample deposited in a collection.

Acknowledgements

Patricia Mergen (Botanical Garden of Meise, Belgium) is thanked for introducing experts who co-authored this paper. We thank Charlotte Duval and Claude Yéprémian of the MNHN (Muséum National d’Histoire Naturelle, Paris, France) for their comments. We acknowledge Axel Hausmann, Matthias F. Geiger and two additional anonymous reviewers for their constructive comments.

This paper was written under COST Action DNAqua-Net. The European COST Action DNAqua-Net (CA 15219 https://dnaqua.net/) is a collaborative network which gathers several hundreds of scientists and water managers, with the objective of developing new genetic tools for bioassessment and monitoring of aquatic ecosystems (Leese et al. 2016). Reference libraries, together with their feasibility and quality, are one of the most important subjects of DNAqua-Net and several meetings highlighted the need to improve the reference data and maintain their quality and reliability. This has been formalised in the discussion of Weigand et al. (2019).

References


https://mbmg.pensoft.net


specimens, an example on parasite hymenoptera (Braconidae). Biological Control 122: 93–100. https://doi.org/10.1016/j.biocontrol.2018.04.001


https://mbmg.pensoft.net


Frédéric Rimet et al.: Standards for building DNA barcode reference libraries for aquatic life


