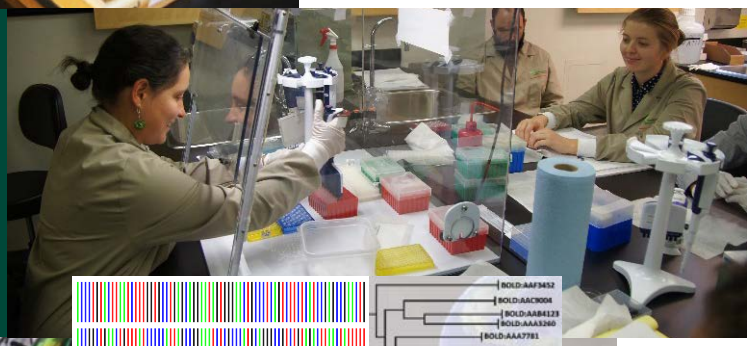
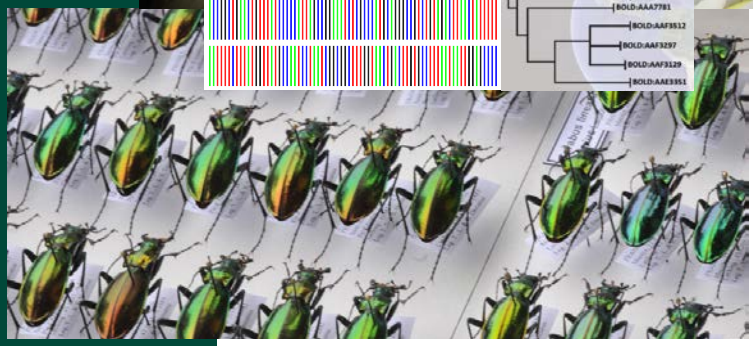
Secretariat of the Convention on Biological Diversity

CBD Technical Series
No. 94

94

The Global Taxonomy Initiative 2020: A Step-by-Step Guide for DNA Barcoding





UN environment programme

Convention on Biological Diversity

gti Global Taxonomy Initiative

Japan Biodiversity Fund 生物多様性日本基金

GTI DNA tech

international BARCODE OF LIFE

CBG

CBD Technical Series No. 94

THE GLOBAL TAXONOMY INITIATIVE 2020: A STEP-BY-STEP GUIDE FOR DNA BARCODING

March 2021

A contribution to the CBD Aichi Biodiversity Targets and beyond.

With support from the Government of Japan
through the Japan Biodiversity Fund.

Published by the Secretariat of the Convention on Biological Diversity
ISBN: 9789292256869 (Print version)
ISBN: 9789292256876 (Web version)

Citation
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 pages.

For further information, please contact:

Secretariat of the Convention on Biological Diversity
413 St. Jacques Street, Suite 800
Montreal, Quebec, Canada H2Y 1N9
Phone: 1(514) 288 2220
Fax: 1(514) 288 6588
E-mail: secretariat@cbd.int
Website: http://www.cbd.int

Photo credits
Carabid beetles by Thibaud Decaëns, all photos courtesy of Centre for Biodiversity Genomics © 2021.

Layout and design: Em Dash Design www.emdashdesign.ca

Foreword

ELIZABETH MARUMA MREMA
Executive Secretary
Convention on Biological Diversity

PAUL HEBERT
Scientific Director and Board Chair
International Barcode of Life (iBOL)

The identification of species is the foundation for evidence-based management of biodiversity and ecosystem functions and services. Access to such information is critical for the effective management, monitoring, reporting, and policy-setting required to achieve the goals of the Convention on Biological Diversity (CBD) while also supporting the post-2020 global biodiversity framework as a stepping-stone towards achieving the 2050 Vision of "Living in harmony with nature".

Through its Global Taxonomy Initiative (GTI), the Secretariat of the CBD has collaborated with the International Barcode of Life (iBOL) consortium to support Parties in expanding their capacity to discover and understand biodiversity. Aided by iBOL, the GTI DNA technology training programme (GTI-DNA-tech) was a five-year effort to address the taxonomic impediment (i.e., lack of scientific experts, taxonomic knowledge and infrastructure) needed to implement the goals of the Convention.

Initial activity focused on providing researchers from developing countries with training on the acquisition and interpretation of DNA barcode records. These individuals subsequently became trainers in their home countries and regions.

Over its lifespan, GTI-DNA-tech enhanced technological and scientific capacity among researchers based in developing, biodiversity-rich countries where it is most needed. It also raised biodiversity awareness by engaging with citizen scientists to extend capacity development. The data-sharing platform developed by iBOL places biodiversity information in the hands of people who need it to make informed decisions. This access is aiding the protection of biodiversity while also strengthening collaborations among a global network of experts in this field.

This issue of the CBD Technical Series targets those who wish to apply DNA barcoding in their own country and learn the laboratory procedures

step-by-step. It provides an overview of the conceptual underpinnings of DNA barcoding and describes the simple workflows and laboratory equipment needed to generate new records. This issue also helps international collaborators to recognize their obligations for access and benefit-sharing surrounding biological specimen collections *in situ* and *ex situ*, in accordance with the three objectives of the Convention and its Protocols. As such, it provides Parties to the Convention with an entree to the application of this rapid, cost-effective technology, furthering implementation of the Convention. It also considers the varied contexts in which DNA barcodes can be applied to advance the conservation and sustainable use of biodiversity.

We believe this issue of the CBD Technical Series will prove a useful resource for all organizations and individuals whose work needs easy access to biological identifications such as those involved in pest management, enforcing bans on trade in endangered species, or environmental impact assessments. We further hope it extends understanding of biodiversity across institutions and countries.

Executive Summary

Most of the multicellular species that share our planet are experiencing population declines, and many are at risk of extinction. To halt this erosion of biodiversity, we need to better manage our interactions with natural ecosystems. The transition towards a greener future requires a global biomonitoring system that tracks the shifts in abundance and distribution of all species. This need cannot be met through biodiversity surveys supported by morphological study, but it can be achieved through DNA barcoding. Aided by advances in DNA sequencing technology and by the development of specialized informatics platforms, DNA barcoding has gained tremendous power over the past 20 years, reflecting the fact that specimen identification and species discovery can be accomplished by analyzing short segments of the genome.

DNA barcoding relies upon the assembly of reference sequence libraries that are linked to specimens that have ideally been identified to a species-level, but this can only be achieved for known species. In practice, specimen identification is often impossible because about 90% of all multicellular species await description, but every specimen can gain placement in higher taxonomic categories. Moreover, so long as it is deposited in a major collection, its taxonomic placement can be refined through time. The DNA barcode workflow begins with the collection of specimens followed by DNA extraction, PCR amplification of the barcode region, and its subsequent sequence analysis. Information on the specimen and its barcode sequence are then deposited in the Barcode of Life Data System (BOLD), the informatics platform developed for this purpose. The DNA barcode reference library on BOLD enables anyone to rapidly ascertain the identity of newly encountered specimens. As a result, DNA barcoding has established itself as a central element in the global biosciences infrastructure and has gained adoption in diverse practical contexts from detecting food fraud to environmental impact assessments.

It is important to note that the country of origin of the organism might have access and benefit-sharing (ABS) obligations attached to its biodiversity sampling (whether *in situ* or in collections). If the user/researcher plans to transfer the specimens across country boders, the international collaborator is responsible for checking and complying with ABS obligations. For the sequence itself, the access and benefit-sharing policy measures around digital sequence information are still being negotiated under the auspices of the CBD.

This contribution to the CBD Technical Series, *GTI 2020: A Step-by-step Guide for DNA Barcoding*, has three goals. It aims firstly to provide background information on the principles underpinning DNA barcoding. Secondly, it discusses the equipment and workflows used to gather and interpret DNA barcodes. Thirdly, it describes both current applications of DNA barcoding and future prospects.

Table of Contents

Introduction

Biodiversity and the Global Taxonomy Initiative

Biological diversity or biodiversity, as defined by the Convention on Biological Diversity (CBD)[1], encompasses genetic variability within species, among species, and among ecosystems. Because species are the primary unit of biodiversity, most studies have focused on them because it is simplest to make quantitative comparisons at this level. Estimates vary, but somewhere between 3 and 100 million species occur on Earth[2], with 8.7 million species (excluding microbes) considered a best estimate. Just 20% of these species have been catalogued since the Linnaean binomial nomenclature system was introduced in the 18th century[3],[4] (**Figure 1**).

The gaps in our inventory of life hinder efforts to understand ecosystems and their functioning limit efforts to take actions which protect species and ecosystems from environmental changes linked to human activities. Recently, the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES)[5] estimated that 1 million animal and plant species are threatened with extinction in this century. This estimate considered previous partial estimates developed by the International Union for the Conservation of Nature (IUCN)[6], as well as Red Lists for various taxa[7]. The actual scale of extinctions will remain uncertain until the number of species is known. The severity of this extinction crisis emphasizes the need to speed the inventory of life. We urgently need the capability to identify all species to understand their interactions as components of the biosphere.

Species identification and discovery has traditionally been based on the examination of morphological characters. This approach is time-consuming and requires expertly trained taxonomists. However, their

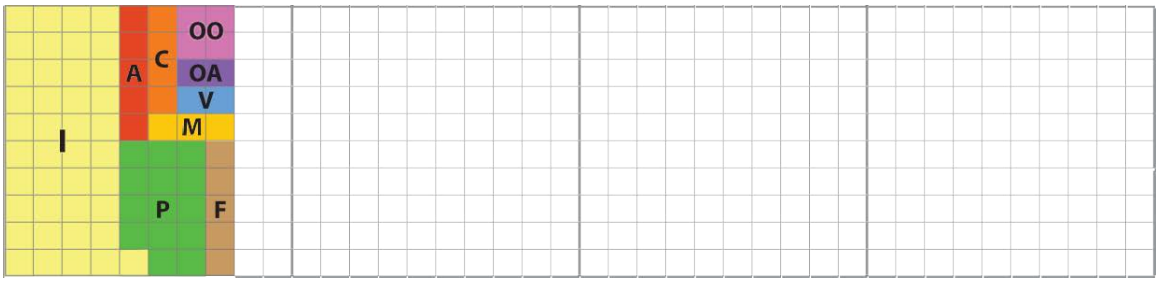


Figure 1. Progress in cataloguing species on Earth. Since the 1750s, about 20% (i.e. the coloured quadrats) of the estimated 8.7 million species of multicellular organisms has been described using morphological approaches. Taxonomic groups: insects (I), crustaceans (C), other arthropods (A), vertebrates (V), molluscs (M), plants (P), fungi (F), other animals (OA), other eukaryotic organisms (OO).

1 CBD (Convention on Biological Diversity). Rio de Janeiro, 5 June 1992.
2 Mora C, Tittensor DP, Adl S, et al. 2011. How many species are there on Earth and in the ocean? *PLoS Biology* 9: e1001127.
3 Linnaeus C. 1753. *Species Plantarum: exhibentes plantas rite cognitas, ad genera relatas, cum differentiis specificis, nominibus trivialibus, synonymis selectis, locis natalibus, secundum systema sexuale digestas* [1st edition]. Laurentius Salvius: Holmiae
4 Linnaeus C. 1758. *Systema Naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Editio decima, reformata [10th revised edition], vol. 1: 824 pp. Laurentius Salvius: Holmiae
5 IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services). 2019. *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES secretariat, Bonn, Germany, 56 pp.
6 https://www.iucn.org/
7 https://ipbes.net/news/how-did-ipbes-estimate-1-million-species-risk-extinction-globalassessment-report

dwindling numbers severely limit our capacity to identify and describe species. The CBD recognized this 'taxonomic impediment,' and, in response, the Parties to the CBD established the Global Taxonomy Initiative (GTI) in 1998[8]. The GTI program has five goals and 18 planned activities[9] , all supporting attainment of the three main objectives of the CBD:

- Conservation of biological diversity
- Sustainable use of its components
- Fair and equitable sharing of benefits arising from the use of genetic resources.

The GTI has supported the development of taxonomic toolkits, has worked to raise taxonomic capacity, and has increased awareness of the important role of taxonomy in informing policy and biodiversity conservation. Once molecular methods, including DNA barcoding, gained acceptance as tools for species identification, GTI incorporated these methods into the scope of its activities.

GTI-DNA-tech and its Role in CBD Implementation

The GTI DNA Technologies Training program (GTI-DNA-tech) was established in 2015 in partnership with the International Barcode of Life Consortium through its Secretariat at the Centre for Biodiversity Genomics at the University of Guelph in Canada with support from the Japan Biodiversity Fund sponsored by the Government of Japan. Since this time, GTI-DNA-tech has provided training to researchers interested in using DNA-based methods to speed species identification in support of CBD implementation at a national level. From 2015-2020, training activities focused on DNA barcoding, an entry point to capacity building in DNA-based species identification in developing countries. These training courses also included brief introductions to newer approaches such as DNA metabarcoding which has gained adoption as the method of choice for large-scale biodiversity inventories and monitoring since 2020.

GTI-DNA-tech has focused on tools for species identification which address the need of Parties to identify threatened, endemic, or invasive alien species, as well as other species of social, economic, cultural, or scientific importance (CBD Articles 7 and 8). By promoting training, access to new technology, and providing technical and scientific guidance, GTI-DNA-tech also addresses the Parties' needs as specified in CBD Articles 12, 16, and 17.

GTI-DNA-tech comprised staged training phases, each building on the experience and lessons from the previous one, culminating with a final training event in 2020. For five years, GTI-DNA-tech provided online and laboratory-based hands-on training at the University of Guelph for researchers (**Figure 2**). Subsequently, these trained researchers trained further researchers on-site in their home countries. GTI-DNA-tech used a 'minimalist approach' requiring only basic infrastructure available in most molecular laboratories. The on-site training had a national or regional focus and addressed national targets of the National Biodiversity Strategy and Action Plans, and the Aichi Biodiversity Targets and Sustainable Development Goals.

GTI-DNA-tech activities during 2015-2020 resulted in a network of approximately 200 trainers across all UN regions. To sustain strong momentum in molecular species identification, trainees will require support from national authorities to continue generating and sharing knowledge that helps integrate biodiversity in all sectors. The Global Taxonomy Initiative Forum at the 14th Conference of the Parties to the CBD (17-29 November 2018, Sharm El-Sheikh,

8 https://www.cbd.int/gti/

9 Secretariat of the Convention on Biological Diversity. 2008. Guide to the Global Taxonomy Initiative. Technical Series No. 30, Montreal, Canada, 156 pp.

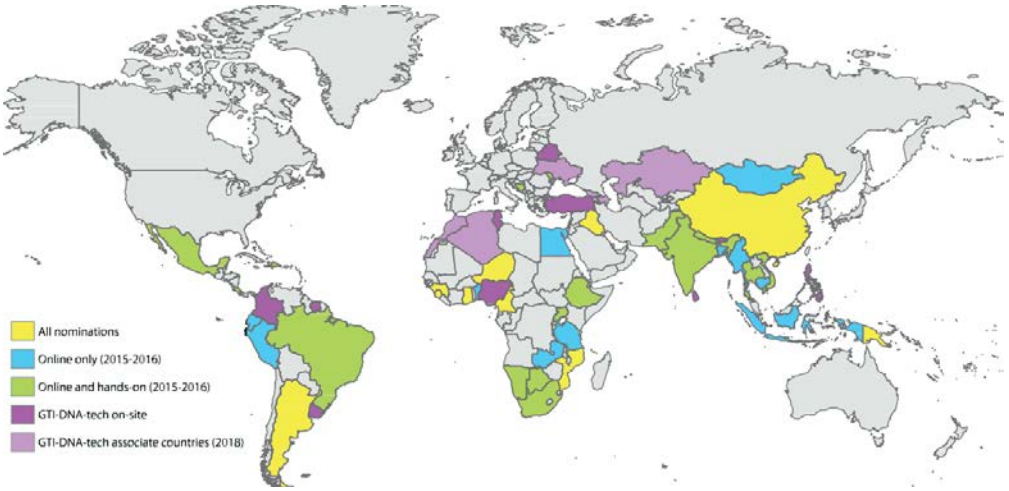


Figure 2. Map of countries participating in GTI-DNA-tech. Home nations for participants in online and hands-on training are indicated by blue and green, respectively. Ten countries that hosted hands-on training events in 2018 are shown in purple while their associated participants are in light purple. Nations that nominated participants that could not be accommodated for training are shown in yellow.

Egypt)[10] highlighted GTI-DNA-tech outcomes. An informative document was also presented to the 23rd Meeting of the Subsidiary Body on Technical, Technological and Scientific Advice (25-29 November 2019, Montreal, Canada)[11].

Purpose of this Guide

This CBD Technical series manual addresses the need for Parties to develop the capacity to rapidly and reliably identify species. Because knowledge of DNA barcoding and the required infrastructure varies across the Parties, it provides a beginner's guide for stakeholders who intend to employ DNA barcoding in their institution and nation to support CBD's objectives.

This guide includes five chapters. The first provides a brief introduction to the concept, history, and components of DNA barcoding. Chapters 2 to 4 provide technical instructions on the three major steps in DNA barcode analysis: sample collection, molecular analysis, and bioinformatics. Chapter 5 outlines applications of DNA barcoding related to CBD implementation.

The guide concludes with a discussion of future directions of DNA technologies for species identification. Initially developed for hands-on training events in Canada and developing countries, this guide has been refined based on experiences at 11 training sessions from 2018 to 2020.

Although this is a CBD technical guide, it has been written for a general audience that lacks background in molecular biology or experience in CBD processes.

This guide provides an overview of DNA barcoding focused on the basic requirements needed to perform molecular species identification. It is not an exhaustive review of available methods and protocols and it is not a replacement for in-depth training. As other molecular techniques (e.g., DNA metabarcoding) evolve, methods will undoubtedly shift to more cost-effective methods for monitoring biodiversity. However, DNA barcoding is the bedrock for molecular species identification and mastering it will provide a solid foundation for involvements in advanced methods.

10 CBD (Convention on Biological Diversity). 2018. CBD/COP/14/INF/12. Available from: https://www.cbd.int/doc/c/c584/aabd/5edf618735fd1b95d2e9732f/cop-14-inf-12-en.pdf

11 CBD (Convention on Biological Diversity). 2018. CBD/SBSTTA/23/INF/18. Available from: https://www.cbd.int/doc/c/6ad1/da5a/ddb684c5c9b0491c89d35872/sbstta-23-inf-18-en.pdf

Chapter 1.

Technical Background

DNA barcoding is a tool that discriminates species by examining sequence variation in standardized gene region that represents a tiny fraction of the entire genome. DNA barcoding gained its name the Universal Product Code, which is employed to identify consumer items. Proposed in 2003 as an identification method for animal species[12], it was subsequently extended to other groups of multicellular life (plants, fungi, protists).

DNA barcoding has now matured into a fast, reliable, and cost-effective tool for species identification and discovery. Through constant refinement of laboratory protocols and technological advances over the past decade, analytical costs have been greatly reduced. The resulting data has allowed DNA barcoding to greatly improve our capacity to catalogue biodiversity and has helped to extend our understanding of species distributions. However, considerable effort is still needed to establish a DNA barcode reference library for all species

Due to its simple workflow, DNA barcoding has gained adoption beyond the taxonomic community (e.g. other scientists, private sector, citizen scientists). It has, in fact, opened the door for anyone interested in biodiversity to become involved in a global effort to collect and share biodiversity data.

The two fundamental principles of DNA barcoding are **standardization** (using the same gene region(s) across large taxonomic groups to facilitate comparisons) and **minimalism** (using the smallest amount of sequence information required for reliable identification of taxa).

DNA barcoding is distinct from DNA taxonomy (species description based only on DNA) and molecular systematics (phylogenetic inference and taxonomic classification based on molecular data). DNA barcoding deals primarily with species identification and has limited utility for lower taxonomic levels (e.g. subspecies, animal breeds, plant varieties). DNA barcodes can aid phylogenetic studies when combined with other molecular markers but have limited phylogenetic signal unless coverage for a taxonomic group is comprehensive.

Standard DNA Barcode Markers

DNA

DNA (deoxyribonucleic acid) is a molecule composed of two nucleotide chains that coil around each other to form a double helix. In most animal cells, about 98% of the DNA is located within the nucleus (nuclear DNA), while the rest is in the mitochondria (mitochondrial DNA; **Figure 3**). In plants, DNA is also present in chloroplasts (plastid DNA). DNA located outside the nucleus is the genetic instructions for its development, functioning, growth, and reproduction. A distinct DNA sequence responsible for the synthesis of a specific product is called a gene. collectively termed organellar DNA. The total DNA of an organism, known as its genome, carries

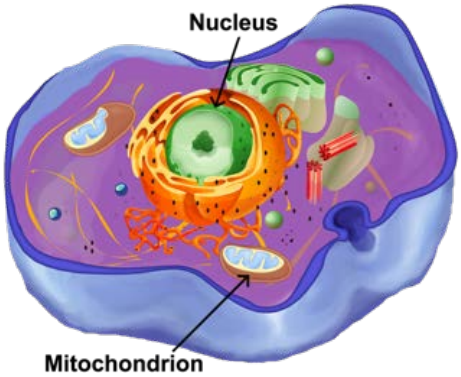


Figure 3. Structure of an animal cell with DNA located in the nucleus and the mitochondria.

12 Hebert PDN, Cywinska A, Ball SL, deWaard JR. 2003. Biological identifications through DNA barcodes. *Proceedings of the Royal Society B* 270: 313–321.

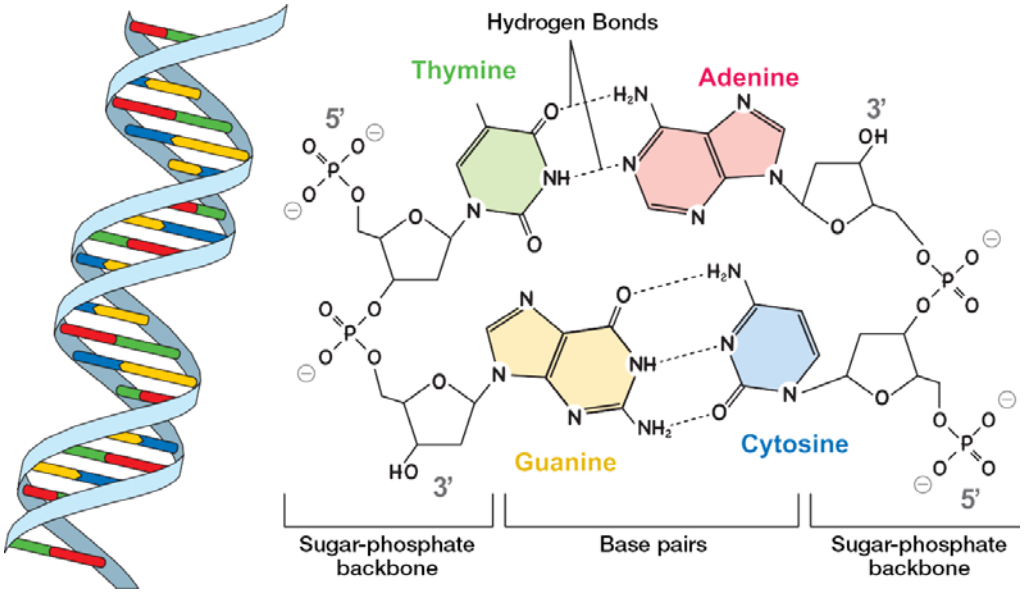


Figure 4. DNA structure: double helix (left) and molecular structure (right).

Each gene is encoded by sequences of four nucleotides, (**A – adenine, T – thymine, C – cytosine, G – guanine)**, each consisting of a nitrogenous base, a sugar (deoxyribose), and a phosphate group (**Figure 4**). Adenine and guanine are classified as purines (two carbon rings and four nitrogen atoms), while thymine and cytosine are pyrimidines (one carbon ring and two nitrogen atoms). The DNA molecule forms a double helix comprised of two strands, resembling a twisted ladder with the backbone made of sugar and phosphate. The strands are joined by hydrogen bonds between complementary nucleotides: A and T are joined by two hydrogen bonds while C and G are joined by three bonds (**Figure 4**). The sequence of nucleotides in a DNA strand is usually read and written in a 5' to 3' orientation. These numbers correspond to the respective terminal carbon atoms in the sugar component at the end of a DNA strand, numbered following a convention in organic chemistry.

Barcode region for animals

The animal barcode region is a 648-base pair (bp) fragment near the 5'-end of the mitochondrial gene cytochrome *c* oxidase subunit I (COI). It was selected due to four key characteristics: i) larger copy number per cell makes it easier to extract and amplify DNA from small amounts of tissue or degraded samples; ii) maternal inheritance and lack of recombination (no exchange of genetic material between maternal and paternal copies of mitochondrial DNA); iii) higher nucleotide substitution rate in mitochondrial DNA results in the rapid accumulation of differences between species; and iv) lack of introns (i.e., non-coding regions within genes which can complicate the comparison of sequences). COI was also chosen as the barcode marker for animals due to its slow mutation rate relative to other mitochondrial genes which aids its recovery via polymerase chain reaction (PCR; **Figure 5**). This gene region can be amplified in many animal species through the PCR (see Chapter 3) using the primer pair LCO1490/HCO2198[13] which are also known as the 'Folmer primers'.

13 Folmer O, Black M, Hoeh W, et al. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology* 3: 294–299.

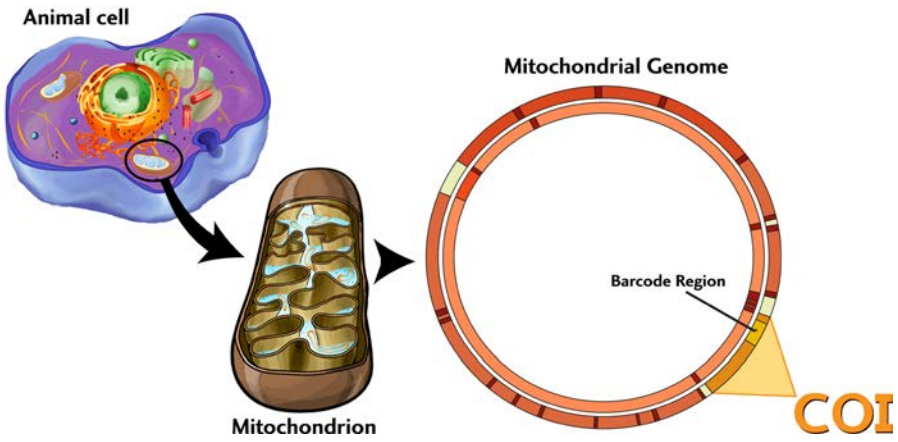


Figure 5. The standard DNA barcode marker for animals is a fragment of the cytochrome *c* oxidase subunit I gene (COI) in the mitochondrial genome.

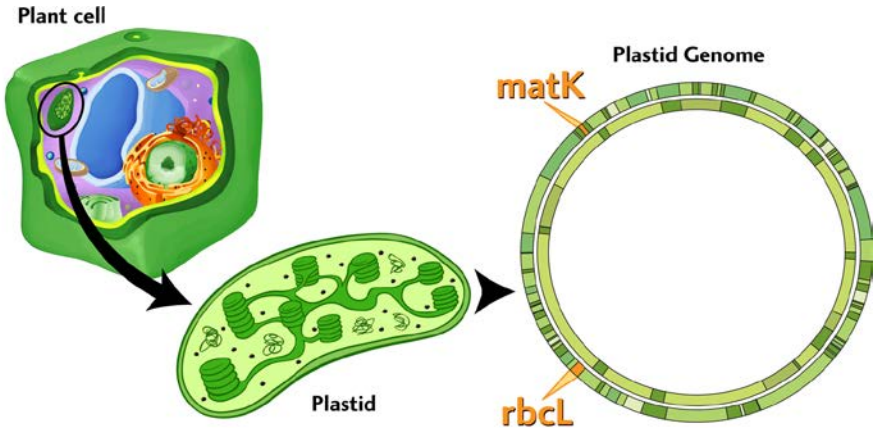


Figure 6. The rbcL and matK genes in the chloroplast genome are the standard markers used in the two-tier DNA barcoding approach for plants.

Barcode region for plants

Standardization, minimalism, and scalability are the core principles of DNA barcoding. While COI meets these criteria for animals, the low rate of molecular evolution in the mitochondrial genomes of plants means there is little divergence between COI in closely related species of plants, ruling out its use as their barcode marker. The search for alternate barcode regions for plants resulted in a recommendation of a two-tiered approach for plant DNA barcoding using chloroplast genes (**Figure 6**):

- First pass analysis using the large-chain subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (rbcL), which can be easily aligned and typically offers genus-level resolution;
- Second pass analysis with maturase K (matK), which can only be aligned among closely related groups of plants but offers improved taxonomic resolution.

Although the combination of rbcL+matK is the accepted standard barcode for land plants[14], other markers are also used[15]. The two most commonly used additional barcode markers for plants are the

14 Hollingsworth PM, Forrest LL, Spouge JL, et al. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106: 12794–12797.

15 Hollingsworth PM, Graham SW, Little DP. 2011. Choosing and using a plant DNA barcode. *PLoS ONE* 6: e19254.

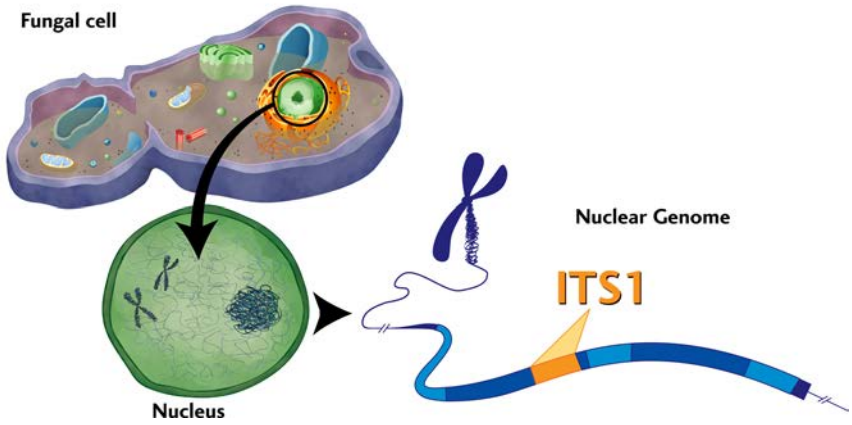


Figure 7. The nuclear internal transcribed spacer (ITS) region is the standard DNA barcode marker for most fungi.

nuclear internal transcribed spacer (ITS), and the non-coding intergenic spacer trnH-psbA in the chloroplast genome[16].

Recent advances in high-throughput sequencing allow the acquisition of large amounts of sequence data at low cost. As a result, genome skimming (shallow sampling of the total genome capturing mainly high-copy fractions such as ribosomal DNA and the mitochondrial and plastid genomes) is emerging as a new tool for identifying plant species[16,17]. However, genome skimming will not be further considered in this guide because it requires specialized training in bioinformatics and more complex laboratory protocols and infrastructure than standard DNA barcoding.

Barcode region for fungi

Fungi, the second-largest kingdom of eukaryotes, are both poorly known and often challenging to identify. The use of COI for fungal identification is complex because their mitochondrial genomes are often rich with long introns, making it difficult to recover a target region through PCR. As a result, the internal transcribed spacer (ITS; **Figure 7**), a non-coding region of the ribosomal cistron in the nuclear genome, was adopted as the standard barcode for fungi[18]. ITS is effective in identifying many fungi. When ITS does not discriminate closely related species, supplemental barcode markers are used[19]. For example, COI is a suitable barcode marker for some fungal taxa (e.g., *Penicillium*)[20] which lack mitochondrial introns.

Barcode region for protists

Protists are a very diverse assemblage containing all eukaryotic organisms that are not classified as animals, plants, or fungi. Their extreme phylogenetic diversity makes it very difficult to find universal DNA barcode markers. Consequently, the protist barcoding community has adopted a nested approach: the V4 region of 18S ribosomal DNA is used as a universal "pre-barcode" sand it is supplanted by additional taxon-specific barcode

16 Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016. From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25: 1423–1428.
17 Dodsworth S. 2015. Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science* 20: 525–527.
18 Schoch CL, Seifert KA, Huhndorf S, et al. 2012. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences* 109: 6241–6246.
19 Lücking R, Aime MC, Robbertse B, et al. 2020. Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding? *IMA Fungus* 11: 1–32.
20 Seifert KA, Samson RA, deWaard JR, et al. 2007. Prospects for fungus identification using CO1 DNA barcodes, with *Penicillium* as a test case. *Proceedings of the National Academy of Sciences* 104: 3901–3906.

markers[21]. The search for effective group-specific barcode regions for some protist taxa is still ongoing.

Applications and Limitations of DNA Barcodes

The digital nature of DNA sequences facilitates automation which is essential for analyzing large datasets. It also escapes the subjectivity inherent in the interpretation of many morphological traits. Furthermore, DNA barcoding can identify specimens of all life stages as well as organismal fragments that lack diagnostic morphological characters, processed products, and even DNA traces in the environment. DNA barcodes can also be used to establish an interim taxonomic system by assigning organisms into operational taxonomic units (OTUs), an effective proxy for species in the absence of an established taxonomic framework. Barcode Index Numbers (BINs)[22] are the most widely used OTU system for animals. They make DNA barcoding very useful for working with poorly known, hyperdiverse and morphologically indistinguishable groups of organisms.

Applications of DNA Barcoding

As a cost-effective, robust approach, DNA barcoding has gained uptake in numerous applications in biodiversity science[23]:

- **Agriculture and forestry**: Identifying and monitoring pests and biological control agents.
- **Human health**: Identifying and monitoring human disease vectors and reservoirs; reconstructing disease transmission pathways; assessing and monitoring vector-borne diseases.
- **Invasive alien species**: Identifying and monitoring alien species that can impact ecosystems and native species; improving early detection and regulatory measures to prevent cross-border transport of alien species.
- **Endangered species**: Improving knowledge of the taxonomy and ecology of endangered species; creating a diagnostic framework to monitor and prevent their illegal harvest and trade.
- **Environmental monitoring**: Supporting the mining (oil, gas, minerals), the conservation (protected areas), natural resource (forestry, fisheries), and agricultural sectors to meet environmental goals and to evaluate the efficiency of resource management, restoration, and mitigation measures.
- **Marketplace surveillance**: Product authentication, detection of food contamination, and substitution (e.g., seafood, meat, nutraceuticals).

Limitations of DNA Barcoding

Aside from resolving taxonomic uncertainties, DNA barcoding can address other research questions such as phylogenetic relationships between species or phylogeographic diversification within species. These applications have sometimes been confounded with DNA barcoding, causing ambiguity in the scope of DNA barcoding. Three categories of limitations exist:

- Conceptual limitation:
  - DNA barcoding is not designed to reconstruct phylogenetic relationships although every barcode region carries some phylogenetic signal.

- Genetic limitations:
  - DNA barcodes may not provide enough resolution to distinguish recently diverged species.
  - Most DNA barcode markers cannot resolve cases of mitochondrial or plastid

21 Pawlowski J, Audic S, Adl S, et al. 2012. CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology* 10: e1001419.
22 Ratnasignham S, Hebert PDN 2013. A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS ONE* 8: e66213.
23 CBD (Convention on Biological Diversity). 2014. UNEP/CBD/SBSTTA/18/INF/20. Available from: https://www.cbd.int/doc/meetings/sbstta/sbstta-18/information/sbstta-18-inf-20-en.pdf

introgression, including ongoing or past hybridization events.
  - Heteroplasmy (i.e., the presence of two or more variants of the barcode region within an individual) may impede the recovery of accurate sequences.
  - Non-functional segments of organellar DNA often occur in the nuclear genome. These nuclear-mitochondrial sequences (NUMTs) and nuclear-plastid sequences (NUPTs) can confound data interpretation if mistaken for the true barcode sequence.

- Methodological limitations:
  - The 'universal' primers for DNA barcoding can fail to amplify the target region in certain groups of organisms.
  - Success in sequence recovery is reduced when DNA is degraded as with old museum specimens or tissue samples exposed to agents that degrade DNA (e.g., high temperatures).

DNA Barcode Data Repositories

The use of DNA barcoding for specimen identification relies on access to openly accessible, well-curated reference databases of DNA barcodes. Following established international practices, sequence data should be made available through deposition in a major online genetic data repository.

The **International Nucleotide Sequence Database Collaboration (INSDC[24])** is the central infrastructure for sharing DNA and RNA sequence data. It includes the DNA Data Bank of Japan (DDBJ[25]), European Nucleotide Archive (ENA[26]) hosted by the European Bioinformatics Institute (EMBL-EBI), and GenBank[27] hosted by the National Centre for Biotechnology Information (NCBI). These three databases (DDBJ, ENA, GenBank) operate separately but share new sequence data each day. Among them, GenBank is the largest, most heavily used repository for DNA sequences.

The broad uptake of DNA barcoding as a tool for species identification has resulted in the generation of extensive barcode sequence data. The unique format and purpose of barcode data required the establishment of a unique platform to serve as both an analytical workbench and data repository.

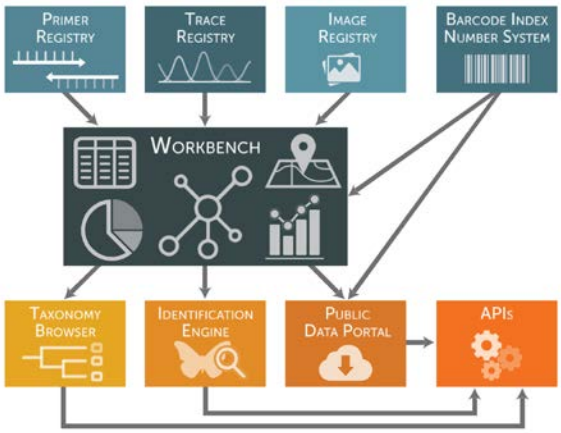


Figure 8. The structure of BOLD. Users can use it to upload, analyze, and publish their barcode data. It also incorporates a unique system for clustering barcode sequences into operational taxonomic units (Barcode Index Number – BIN)[28]. It offers a public database that can be queried/downloaded, an identification engine for unknown sequences, and an API interface for automatic and programmable user requests to BOLD.

These needs spurred development of the **Barcode of Life Data System (BOLD)** platform[29,30] (**Figure 8**) to host and analyze DNA barcode sequence information, associated raw

24 http://www.insdc.org/
25 http://www.ddbj.nig.ac.jp/
26 https://www.ebi.ac.uk/ena/
27 https://www.ncbi.nlm.nih.gov/genbank/
28 Ratnasingham S, Hebert PDN. 2013. A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS ONE* 8: e66213.
29 http://boldsystems.org
30 Ratnasingham S, Hebert PDN. 2007. BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* 7: 355–364.

data, provenance details, images, and taxonomic annotations related to the organisms from which the barcode sequences originate. Its architecture incorporates modules designed to store, organize, visualize, review, curate, analyze, and share DNA barcode datasets to facilitate collaborative research and application. Moreover, it is linked to the INSDC, so BOLD users can submit their sequence records directly to GenBank, and BOLD can perform data mining for specific markers (mainly barcode markers).

DNA Barcoding Workflow: General Overview

The DNA barcoding workflow consists of three main components: (i) specimen collection and management, (ii) molecular analysis, and (iii) informatics (**Figure 9**).

Depending on the project's purpose – reference library construction versus library use for applications – these three components can vary. For instance, for the construction of reference libraries, it is critical to store each barcoded specimen as a voucher in a public institution and to upload all related metadata to BOLD. By contrast, in many applications (e.g., identification of a fish fillet), work simply focuses on acquiring a sequence from the sample and querying the resultant sequence against the online repositories (GenBank and BOLD). The next sections of this guide focus on the construction of reference libraries, but also provide context for the simpler workflows employed in many applications of DNA barcoding.

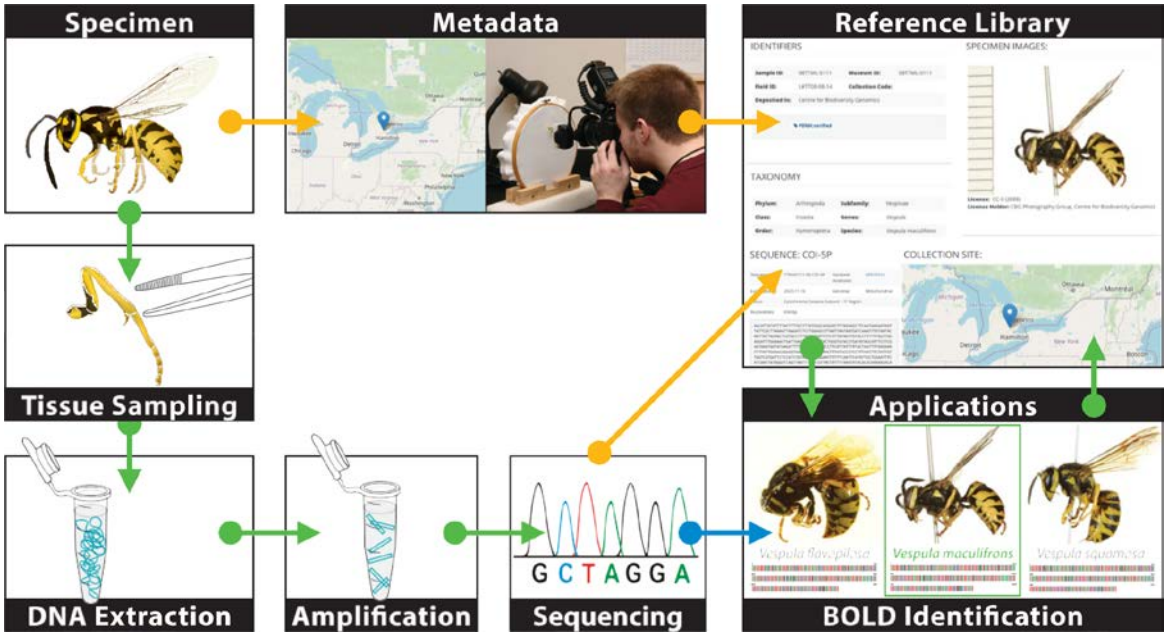


Figure 9. The DNA barcoding workflow from specimen collection to uploading sequences and metadata. Yellow arrows indicate points of data upload to BOLD.

Chapter 2.

Collection Management

For centuries, specimens collected during expeditions and preserved in natural history collections have formed the foundation for efforts to inventory biodiversity. These collections provide a wealth of information about species and their distribution. Today, they also provide important baseline data to detect shifts caused by anthropogenic disruptions such as climate change or human-mediated dispersal of organisms outside their native range.

It is crucial that specimens associated with sequence records in DNA barcode reference libraries are properly stored in natural history collections and available on request for further examination. This chapter focuses on collecting and maintaining physical specimens as part of the global effort to build comprehensive barcode reference libraries hosted on BOLD (see BOLD handbook[31] for information on how to upload data to BOLD).

The Society for the Preservation of Natural History Collections (SPNHC)[32] has compiled a variety of materials regarding best practices for establishing and maintaining various types of natural history collections (insects, vertebrates, herbaria, etc.). Topics include dry and fluid collections, collection restoration, labelling and digitization of specimens, and information on access and benefit-sharing. For anyone planning to establish a new collection, especially at the regional or national level with a taxonomic or geographic scope, the SPNHC website[31] is the ideal place to obtain information.

Rather than investing in a new collection that might not be sustainable in the long term, it is usually advisable to deposit the voucher specimens from DNA barcoding studies in a well-established repository with the expertise and infrastructure required to store biological material and supported by adequate funding.

State-of-the-art DNA barcoding in a natural history collection environment includes best practices to ensure the highest data quality (**Figure 10**). Collecting and storing material in

Planning the collection

A few points to be considered before starting to collect material for DNA barcoding:

1) Is specimen preservation necessary? Will the specimens become part of a reference library or will they be discarded after DNA extraction?
2) Is there a need to establish a new collection or can specimens be stored in an existing national/local natural history collection?
3) If a new collection is to be established:
   a. What are the national rules and regulations?
   b. What are the best practices for building and maintaining a collection?
   c. What infrastructure is needed?
   d. What equipment and consumables are needed?
   e. What human resources are needed?
4) What is the sustainability plan for long-term persistence of the collection (at a scale of many decades)?
5) Is there any outreach strategy to maximize the value of the collection beyond purely storing biological material for research purposes?

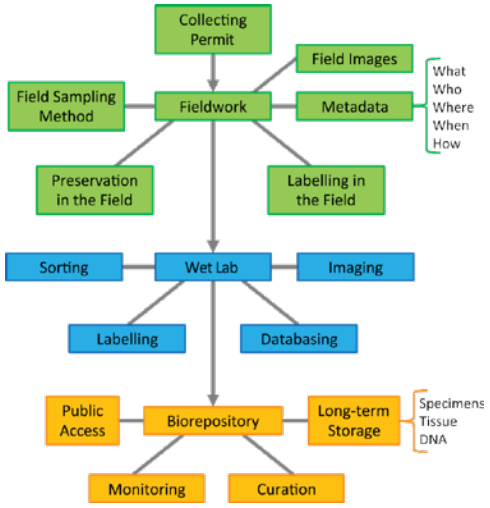


Figure 10. Collection management: planning schema for specimen acquisition, processing, and storage. Green: Fieldwork and specimen collection. Blue: Laboratory organization of specimens for molecular processing. Yellow: Long-term storage and curation of biological specimens, tissue, and DNA including public access.

31 http://boldsystems.org/libhtml_v3/static/BOLD4_Documentation_Draft1.pdf
32 https://spnhc.org/

a 'DNA-friendly' manner ensures the availability of tissue or DNA extracts for future molecular analyses, such as the subsequent sequencing of additional DNA fragments or entire genomes of the preserved specimens.

Specimen Collection

Fieldwork should always involve careful planning and organization ahead of time to limit errors in the field. Implementation requires administrative (e.g., permits and authorization), physical (e.g., equipment, travel logistics), and financial resources that, again, need to be carefully considered. It is important to clarify some key terms commonly found in the literature (see text box below). More details about collection terminology in biodiversity science can be found online[33].

Collecting permits

Before organizing fieldwork, it is mandatory to acquire all necessary permits required by local and national authorities. Every country has a set of regulations regarding work with biological organisms (collecting, handling, exporting to

Collecting Activities

Collecting Effort	Activity with an overarching goal (e.g., surveying specific organisms within a territory over a certain period), part of an institutional or collaborative project or program. *Example: Expedition to a geographic location or research program that involves recurring collecting activities in a certain area.*
Collecting Event	Targeted sampling activity focused on a taxonomic group, a particular locality, or a short time period. Multiple events contribute to one collecting effort. *Example: One Malaise trap sampled at one time; a pan trap and a pitfall trap placed in the same location at the same time would represent two collecting events.*

Collection Objects

Lot	Bulk sample resulting from a collecting event, stored in one container. It consists of multiple unidentified biological organisms. *Example: Bulk sample from a Malaise trap. Each lot needs to be preserved separately and labelled appropriately.*
Specimen	A single individual, either separately collected in the field or removed from a bulk sample (lot). In the latter case, it is crucial to retain a link to the originating lot. Also referred to as 'collection voucher' (or 'voucher specimen' or simply 'voucher' in this guide).
Tissue sample	A portion of a specimen (usually a piece of DNA-rich tissue) preserved for molecular analysis. For microscopic organisms, whole individuals may be consumptively analyzed. The exoskeletons of hard-bodied organisms such as many insects can be recovered after lysis and stored as vouchers. For larger organisms (e.g., vertebrates), one or more tissue samples from a single individual can be stored in a tissue collection. Only a small portion of a sample is used for DNA barcoding. *Example: Insect leg.*

Note: An important distinction should be made between specimen and species: a **specimen** is a physical entity (biological individual) while a **species** is an operational unit used to group specimens based on a set of criteria.

33 Walls RL, Deck J, Guralnick R, et al. 2014. Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies. *PLoS ONE* 9: e89606.

another country etc.) that need to be respected. These regulations may even vary within a country, e.g., between provinces or states. Additionally, many rare species are regulated by international agreements such as the Convention on International Trade in Endangered Species in Wild Fauna and Flora (CITES)[34]. Collecting and exporting/importing representatives of these species requires a set of additional permits.

The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity[35] entered into force in 2014. 129 Parties to the CBD ratified the Protocol as of February 2021. Each Party has taken appropriate legislative, administrative, or policy measures regarding the utilization of genetic resources[36]. Collecting biological organisms and analyzing their DNA sequences, for example, is especially important when organisms are collected in biodiversity-rich countries and exported to other countries for research and other purposes.

Sampling methods

A variety of sampling protocols are employed to collect organisms (**Figure 11**). They are usually categorized into active and passive methods. Active methods require continued trap handling by a collector, while passive ones rely on sampling devices that are left unsupervised in the field for a period of time. The latter can be easily standardized and usually come with a better cost-benefit ratio. A few methods commonly used to collect organisms for DNA barcoding or other molecular studies are listed in **Table 1**.

Some research should be done before fieldwork to gather all the logistical details necessary for the target taxon (sampling method, gear needed, chemicals to kill and store organisms, jars and vials, labelling system, transportation from the field to the lab, etc.).

In some cases, especially for large specimens where it is impossible to collect the entire organism, the preservation of one or few tissue samples, augmented with photographs of the live specimen, is recommended. The most important

Table 1. Common methods used to collect organisms for DNA barcoding studies.

TAXON	SAMPLING METHODS
Terrestrial insects, other arthropods	Malaise traps, pan traps, pitfall traps, sifters, sweep nets, UV light sheets, light traps
Aquatic insects, macroinvertebrates	Kick nets, plankton nets, bottle traps, underwater light traps
Marine invertebrates	Plankton nets, benthic trawls and grabs, diving, Autonomous Reef Monitoring Structures (ARMS)
Fish	Gill nets, electrofishing
Birds	Mist nets
Amphibians, reptiles, mammals	Pitfall traps, Sherman traps, nets, hand collecting
Plants, lichens, fungi	Hand collecting
Seaweeds	SCUBA diving, hand collecting

34 https://www.cites.org/
35 https://www.cbd.int/abs/
36 See https://absch.cbd.int/ for details.

Figure 11. Collecting invertebrates for DNA barcoding in the Canadian Arctic. Left: Sifting leaf litter for spiders, mites, and beetles. Right: Dredging for benthic marine invertebrates.

aspect to be considered is a DNA-friendly way of collecting and storing organisms/tissue in the field[37].

If fieldwork extends over many days, it is important to establish a daily routine:

- Collect specimens
- Gather additional metadata at site (see section below)
- Pre-process specimens once back at the main base of operations.

For instance, when collecting aquatic invertebrates (freshwater or marine), these are brought back from the field in water or ethanol. Particles such as sediment or plant debris should be removed, and specimens placed in fresh ethanol. Replacing the initial batch of ethanol with fresh ethanol after 12-24 hours improves DNA preservation. To further aid DNA preservation, samples should be kept cool even under field conditions. For botanical work, each plant should be pressed between newspaper sheets or transferred into silica back at the base.

Ethanol (96-100%) is the most commonly used chemical for killing and preservation of organisms in the field. Invertebrates killed directly in ethanol will lose colour, and the pigments will leech into the ethanol, which is why it needs to be replaced after 12-24 hours. Also, the use of cold ethanol in the field increases the chances of recovering good quality DNA[38].

DNA-friendly killing/preservation in the field

- Non-chemical methods: freezing, drying
- Supersaturated salt solution
- Cold ethanol: most invertebrates (organisms will lose colour)
- Chloroform, cyanide, ammonia: insects
- Isoflurane, carbon dioxide (vertebrates)
- RNAlater: smaller organisms (organisms will keep their colour) or tissue (expensive chemical)

To be avoided:

- Formalin (degrades DNA)
- Propylene glycol

37 Gonzalez M. A., Arenas-Castro H. (Eds). 2017. Recolección de tejidos biológicos para análisis genéticos. Instituto de Investigación de Recursos Biológicos Alexander von Humboldt. Bogotá, D. C., Colombia. 33 pp.

38 Prosser S, Martínez-Arce A, Elías-Gutiérrez M. 2013. A new set of primers for COI amplification from freshwater microcrustaceans. *Molecular Ecology Resources* 13: 1151–1155.

Recording Metadata

During fieldwork, some crucial data need to be collected on site because recalling every detail after returning from the field is difficult, especially if multiple samples were collected throughout the day. These data include sampling date, collector name, locality, GPS coordinates, type of habitat, sampling method, and any other relevant details. Typically, collecting details are recorded on site in a notebook and later inserted into a spreadsheet on a computer.

Tips and tricks

Field conditions can be challenging for multiple reasons. The use of special paper resistant to rough weather conditions, such as Rite-in-the-Rain paper, and pencils help protect the notes taken while conducting fieldwork.

Place proper **labels** in collecting jars and vials at the collecting site in order to reliably connect specimens to collection metadata throughout the barcoding workflow. A clear and non-duplicating alpha-numerical coding system and pre-printed labels will prevent errors in the field.

Photographs of the field site, collectors manipulating gear, and collected material are always valuable for presentations, reports, and outreach material.

Processing Samples after Fieldwork

For a typical DNA barcoding workflow, specimens destined for analysis need to be sorted from bulk samples into individual specimens, placed in individual vials with fresh 96-100% ethanol (or pinned – insects, pressed – plants), labelled, databased, and imaged, before tissue sampling and DNA extraction. After tissue sampling, specimens should be deposited as 'vouchers' in a publicly-accessible collection (**Figure 12**). Some specimens are collected and labelled individually directly in the field (e.g., bigger invertebrates such as crabs) or tissue sampled in the field (e.g., blood or hair sample from a mammal, or leg segment from an endangered butterfly before its release). Regardless of the sampling protocol, the preparation of specimens/samples for molecular processing and subsequent long-term storage requires proper labelling, preservation, and databasing.

Labelling

Proper labelling is a critical factor when handling biological specimens for research and subsequent long-term storage. As mentioned above, each specimen code must use a unique combination of letters and numbers to avoid duplication in the final collection. Moreover, the type of paper and ink used for labelling, especially for fluid collections where labels are often submerged in ethanol

Figure 12. Examples of voucher specimens preserved in natural history collections. Left to right: vertebrate specimen (bird study skin), entomological specimen (pinned beetle), plant specimen (herbarium sheet). Each specimen has a label which includes a globally unique voucher number that makes it possible to track each specimen through the analytical process and to link it with the corresponding data records and DNA sequences.

for decades at a time, is key. SPHNC provides documentation on this topic[39] that should be consulted in the planning phase.

Tips and tricks

For samples stored in ethanol, there should be a label inside the vial and a duplicate on its outside surface if labels are printed (some printer ink is affected by ethanol). If labels are handwritten, the use of an alcohol-proof pen or a pencil is essential.

Imaging

Digital images provide an independent way of verifying a voucher specimen's taxonomic identity, especially when it is lost or inaccessible. Specimen images, also called **e-vouchers**, form a central component of the reference DNA barcode database. Although it may not be possible to capture all morphological diagnostic features with a single image, it often provides sufficient information to resolve discordances between the assigned taxonomy of the imaged specimen and the identification inferred from its DNA barcode sequence. BOLD supports the storage of multiple images per specimen for specimens processed through DNA barcoding while other repositories (e.g., MorphBank[40]) store general specimen images used in research collaborations and education.

It is ideal to image specimens before molecular analysis to capture relatively intact morphology, especially if whole specimens are used for DNA extraction. In some cases, specimens require special preparation to generate diagnostically meaningful images (e.g., slide mounting, which can only be done after tissue sampling or whole specimen tissue lysis). In these situations, imaging must be done after barcode analysis.

Macro versus microphotography

Depending on the size and nature of specimens and mode of preservation (dry versus fluid), either a camera with a macro lens (**Figure 13**) or a camera attached to a microscope (microphotography) are used.

Macro photography uses less expensive equipment and comes with greater versatility in the studio setup, higher optical quality, and higher throughput. Although most 'consumer-grade' cameras offer a macro function, it is better to use single-lens reflex (SLR) cameras with interchangeable lenses. Most macro lenses can capture full-frame images of specimens 10 mm and larger. However, high-resolution images of smaller specimens (e.g., >2 mm) can also be imaged if the original image is cropped. When working with large vouchers (e.g., vertebrate or herbarium specimens) wide-angle lenses can be used.

Microphotography is essential for small specimens (<2 mm), particularly if they are in ethanol. Working with such specimens is usually

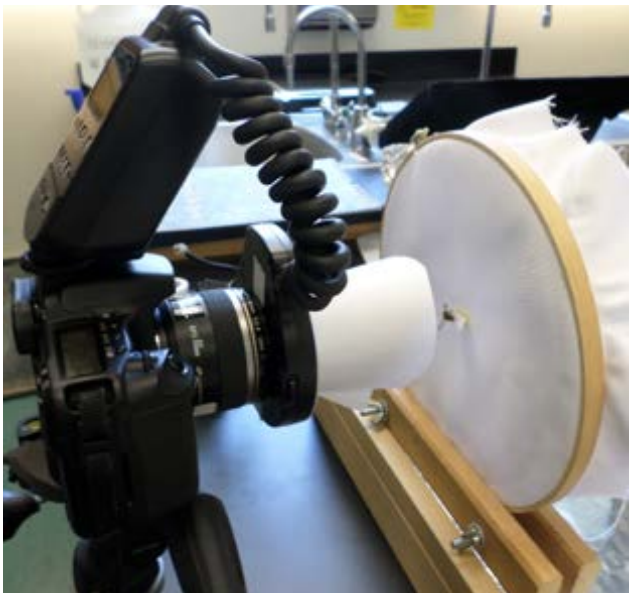


Figure 13. Example of a macro photography setup for pinned insects. The camera is mounted on a tripod; a ring macro flash is attached to a 60 mm macro lens. The cone of white paper in front of the flash acts as a diffusor that minimizes glare. The specimen is pinned onto a double-layered piece of white fabric. There is an off-camera flash behind the fabric that is synchronized with the on-camera ring flash. When triggered, it overexposes the fabric, creating the perception of a pure-white background.

39 https://spnhc.biowikifarm.net/wiki/Labeling_Natural_History_Collections#Paper
40 https://www.morphbank.net/

more time-consuming. Microscopes are typically stationary and often more expensive than cameras. Many require specific cameras and proprietary software to capture and process images. If the specimens exceed 0.5 mm, medium to high magnification stereoscopic ('dissecting') microscopes are preferred over compound microscopes because they allow direct manipulation of the specimen. This enables the combination of imaging and tissue sampling in a single step of the workflow (**Figure 14**).

Processing software

Various software can be used for the post-processing of images before they are uploaded to BOLD (or another online platform). FastStone Image Viewer[41] is a versatile option which is free for personal and educational use. It supports basic image adjustment, cropping, batch resizing, and batch renaming of files.

There are multiple online sources[42] for detailed information about the best protocols for imaging biological specimens.

Figure 14. Examples of a microphotograph (A) and macrophotograph (B). A: image of a slide-mounted flea taken using a digital camera mounted on a dissecting microscope. B: photograph of a fluid-preserved frog specimen taken using an SLR camera with a macro lens.

Tips and tricks

When imaging specimens that will become part of DNA barcode libraries and uploaded to BOLD, time and effort can be saved by naming each image file with its corresponding Sample ID or Process ID of the voucher specimen.

- Sample ID: unique code given to each specimen by the user
- Process ID: unique code given to each specimen by BOLD

Biorepositories

Long-term preservation of biological specimens is a critical component of DNA barcoding that must be considered before starting a project. Ideally, all barcoded specimens are stored in a natural history collection within their country of origin to allow local researchers to extend local knowledge of biodiversity.

Large natural history collections are usually divided taxonomically (e.g., insect collections, vertebrate collections, herbaria), by the type of preservation (fluid collections versus dry collections), or sometimes by geography (local/regional collections). Traditionally, fluid collections contained specimens (largely marine and freshwater invertebrates) collected in bulk and stored immediately in formalin. Upon return

41 https://www.faststone.org/
42 https://www.ala.org.au/wp-content/uploads/2011/10/BK-Guidance-on-Photographing-specimens_FINAL.pdf

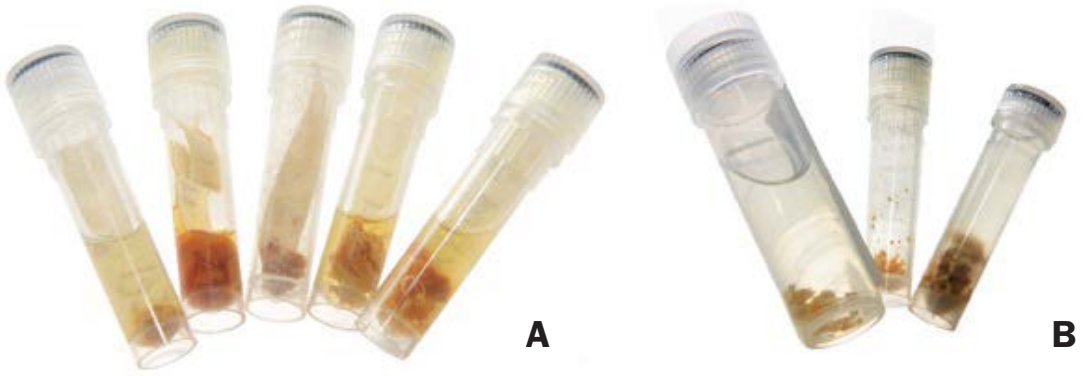


Figure 15. Examples of ethanol-preserved vertebrate muscle samples with (A) low and (B) high likelihood of DNA barcode recovery. If the sample volume is large relative to ethanol, if the sample has not been separated into small pieces, or if the ethanol is coloured and partially evaporated, chances of recovering DNA barcode sequences are lower.

to the museum, they were sorted and placed in 70% ethanol (sometimes mixed with glycerin). Formalin preserves colours, and also keeps invertebrate bodies flexible (arthropods such as small crustaceans become brittle in ethanol, making them more difficult to manipulate for morphological identification). However, its use closes the door for molecular analysis and it also carries human health risks.

With the emergence of DNA technologies, researchers recognized the importance of using DNA-friendly methods to collect and preserve biological material. As a result, an increasing number of collections are now storing at least subsamples of tissue under conditions suitable for genetic analyses. At the same time, the remainder of a voucher is kept separately using traditional preservation methods.

Ethanol preservation

Despite its wide usage, ethanol does not guarantee the preservation of high-quality, high molecular weight DNA due to several factors:

- Ethanol quality: acidity and additives may damage DNA
- Ethanol concentration: elevated water content leads to DNA hydrolysis
- Sample/ethanol ratio: excessive tissue lowers ethanol concentration and increases the concentration of autolytic enzymes (ratio recommended = 1 sample: 3 ethanol; **Figure 15**)
- Storage temperature: high temperatures lead to DNA degradation
- Light: light can lead to DNA photolysis
- Ethanol evaporation: evaporation increases water concentration, accelerating hydrolysis.
- DNA degradation is slowed if samples are stored in cool conditions, ideally below 0°C.

Dry (desiccated) preservation

Another useful method for DNA preservation is desiccation. As it also preserves most morphological features, it is often used in herbaria and insect collections (**Figure 16**).

The following factors need to be considered to ensure DNA degradation is slowed:

- Drying conditions (how long and how well)
- Pre-treatment procedures (skin tanning, insect relaxing) may have adverse effects
- Ambient humidity should be low (one of the most important factors affecting collections in tropical countries)
- Storage temperature should be as low as possible
- Exposure to light should be minimized
- Avoid use of fumigants and preservatives during preparation or storage

Figure 16. Examples of dry-preserved specimens: dried butterfly removed from a storage envelope (A) and plant leaf sample in a bag with silica gel as desiccant (B). Whole insects can be preserved in envelopes or (more commonly) pinned. Plants are typically dry preserved on herbarium sheets with tissue samples destined for molecular analyses often stored separately in sealed bags containing silica gel to provide better protection against fluctuations in humidity.

BOLD Data Submission – Reference Library

Voucher details (taxonomy, collection details, and biological details of specimens) need to be uploaded to BOLD into a previously created project. Once this step is completed, a photograph of each voucher can be uploaded. Although this photograph may not display enough morphological characters for a species-level identification, images are critical for data validation (e.g., detecting cross-contamination between specimens). Instructions for data submission are provided in the BOLD Handbook.

Chapter 3.

Molecular Analysis

This chapter provides basic information on the protocols used to process specimens in a molecular laboratory. It describes both the basic infrastructure required to carry out work and the standard protocol for each processing step. It is possible to outsource the molecular pipeline or some of its steps to a dedicated molecular facility.

Setting up a small laboratory for in-house DNA barcoding requires staff with some molecular experience (e.g., pipetting, sample and reagent handling, contamination avoidance, equipment handling, background in chemistry and molecular biology).

Molecular Laboratory Set-up

A molecular laboratory needs to be in a closed physical space separated from offices and common areas. Ideally, several rooms are available, each dedicated to a specific processing step to prevent contamination (e.g., reagent preparation and storage, sample preparation, molecular processing, electrophoresis). However, work can be conducted in a single room by working with care. To prevent contamination, it is important to have at least some physical and **spatial separation** of equipment and consumables, so the **workflow is unidirectional**.

Table 2 lists the basic infrastructure needed to conduct DNA extraction, PCR amplification, and verification of amplification success by gel electrophoresis. The final step, DNA sequencing, requires DNA sequencers that are not present in many organizations. Outsourcing DNA sequencing to commercial providers is a cost-effective solution.

As technology is rapidly evolving, new equipment emerges on as regular basis. For instance, the Bento Lab[43] provides a basic setup for processing of small numbers of samples from DNA extraction to Gel electrophoresis.

A tracking system for all laboratory steps is crucial for successful DNA barcoding. Larger organizations always employ a Laboratory Information Management System (LIMS) to track every step in the workflow and to document the protocol used by their technicians. Alternatively, spreadsheets can be used for tracking purposes, although they are more labour-intensive.

Good Laboratory Practices

Good laboratory safety practices should be followed to prevent risks to staff or the contamination of samples.

General Guidelines

- No food or drink inside the laboratory
- All working stations properly cleaned before and after use (use 70% ethanol and bleach solution, UV light if available)
- Waste (general, recyclable, hazardous) discarded as appropriate
- All necessary equipment and reagents prepared before starting lab work

Laboratory Structure

- Separate workstations with separate pipettes for different steps (reagent preparation, DNA extraction, amplification, gel electrophoresis)
- No bidirectional workflow (especially pre- and post-PCR: no handling of amplified PCR products at or near DNA extraction / PCR preparation workstations)
- If ethidium bromide (known mutagen) is used for visualization of amplicons during electrophoresis, the workstation is a 'contaminated and hazardous area'. Extra care should be taken (separate lab coat, separate pipettes, workspace isolated from rest of the lab). Waste produced (used gels, pipette tips, gloves) should be viewed as hazardous. There are alternative, non-toxic options available
- Chemicals should be labelled and stored according to manufacturer's instructions (incompatible chemicals such as acids and bases should not be stored together). The date

43 https://www.bento.bio/

Table 2. Basic infrastructure for a DNA barcoding laboratory

ITEM	NOTES
Biosafety cabinet	For PCR setup if no separate room is available.
Fume hood	Needed if using volatile chemicals.
Centrifuge	A refrigerated centrifuge that can handle high speed; mini centrifuges are useful for short spins.
Incubator	Can be replaced with a heated water bath.
Thermocycler	Fast ramping thermocyclers produce faster results.
Autoclave	Needed to sterilize tubes and pipette tips (unless new tubes and pipette tips are used for each analysis).
Fridge	If only one fridge is available, sealable plastic containers should be used to separate reagents from samples to avoid contamination.
Freezer	If only one freezer is available, sealable plastic containers should be used to separate reagents from samples or DNA extracts to avoid contamination.
Cooling block or ice maker	To keep reagents cold during PCR setup (unless the polymerase is stable at room temperature).
Electrophoresis set-up	Electrophoresis chambers have different sizes depending on the number of samples that they can run
Gel imaging and documentation system	To visualize and image gels under UV light (or blue LED light) after electrophoresis.
Spectrophotometer or Fluorometer	DNA quantification before PCR. Although good practice, it is not essential.
Vortex mixer	For tissue preparation and reagent preparation.
Magnetic stirrer with heater	Needed if reagents are made in-house.
Microwave	To melt agarose for gel electrophoresis unless precast gels are used.
Pipettes	Multiple pipettes needed (0.2-10 µL; 2-20 µL; 20-200 µL, 100-1,000 µL).
Plasticware	Tubes of different volumes (0.5 mL to 50 mL), tips (different volumes; filter tips are recommended to prevent contamination).
Glassware	Needed if reagents are made in house and for gel preparation.
Gloves	Sterile powder-free gloves, different sizes.
Other tools	Permanent markers, notebooks, autoclave tape, pens, forceps, spatulas, mortar and pestle (for plants).
Cleaning supplies	70% ethanol and bleach solution.

the container was opened or when the reagent was made should be written on the label. Flammable reagents (e.g., ethanol) should be kept in a flammable storage cabinet
- Material Safety Data Sheets (MSDS) received with the chemicals should be stored in a visible place – in case of contact with a potentially hazardous chemical, the MSDS should be consulted to learn the best course of action

Personnel Safety

- All personnel should receive safety training and should be skilled in laboratory practices such as proper handling of equipment and chemicals
- Lab coats, gloves, and closed toe shoes must be worn while working in the laboratory to avoid direct contact with chemicals as well as contamination of samples or reagents
- Safety glasses and face masks should be used when handling volatile chemicals
- Extra care should be taken at the gel electrophoresis station if ethidium bromide is used
- First-aid kits and fire extinguishers should be stored in a visible location
- Phone number of the emergency contact person should be clearly visible

Tissue Sampling

The quantity of tissue needed for DNA extraction depends on the protocol used; each DNA extraction kit includes detailed instructions. **Table 3** lists recommended tissue quantities for various taxonomic groups.

Items needed for tissue sampling:

- Forceps (smooth tipped, not ribbed) or blade
- Ethanol/gas burner and lighter
- Ethanol in small jars to dip the forceps tips before flaming

Table 3. Common tissue quantities used for DNA extraction of different taxa

TAXON	TISSUE	QUANTITY
Insects, other arthropods (small)	Whole leg, antenna	~ 5 mm length
Insects, other arthropods (large)	Partial leg (tibia or femur)	~ 2 mm length
Other invertebrates	Muscle	~ 1 mm^3
Minute invertebrates	Whole specimen	< 3 mm length
Fish	Fin clips, muscle, eyeball	~ 1 mm^3
Amphibians	Toe clippings, swabs	~ 1 mm^3
Reptiles	Tail, blood, swabs	~ 1 mm^3
Birds	Feathers, muscle, blood	~ 1 mm^3
Mammals	Muscle, tail, ear punches, blood, swabs	~ 1 mm^3
Plants	Leaf, stem	~ 50 mm^2
Fungi	Fruiting body or mycelium	~ 50 mm^2

Notes:
- Excessive tissue can inhibit DNA extraction and amplification.
- Avoid sampling near the digestive tract due to potential bacterial contamination.
- Forceps or blades used to cut tissue must be sterilized after each sample.
- Each tube with a tissue sample needs to be well labelled (use a permanent marker).

- Sterile centrifuge tubes (usually 1.5 or 2 mL) to receive the tissue
- Rack for tubes
- Permanent markers to label tubes
- For plants: Mortar and pestle to break leaf tissue; spatula to transfer the tissue into the tube – mortar, pestle and spatula need to be sterilized between samples
- Petri dish to place the forceps while not being used
- Tissue paper to wipe excess tissue off the forceps before sterilization
- Gloves

Forceps can be sterilized in two ways (**Figure 17**):

- Flame sterilization (especially for DNA-poor tissues such as insect legs and plant leaves) – dip the tip of the forceps in ethanol, pass it through flame and wait a few seconds for the flame to disappear and the forceps to cool off before harvesting the next sample.
- Bleach/ELIMINase (especially for DNA-rich tissues like vertebrate muscle) – dip the tip of the forceps into a jar with bleach/ELIMINase, then consecutively into three jars with distilled water to remove any trace of the chemical.

DNA Extraction

DNA extraction is the process of isolating DNA from other cellular components. The source of DNA can vary from whole specimens to fragments of skin, muscles, feathers, organs, gut contents, feces, seeds, pollen, and even body swabs or cells shed into the environment. Whether the target DNA fragment is nuclear, mitochondrial or plastid, total genomic DNA is usually extracted in a process which includes two main steps: releasing DNA from the cell (by disrupting nuclear, organellar, and cell membranes through lysis), and separating DNA from the other compounds (by isolation). The DNA extraction method used depends on the organism, as well as source, age, and size of the sample. For example, plants have strong cell walls while animal cells do not so plant tissues require the use of additional chemicals and modified protocols for lysis (e.g., tissue ground with mortar and pestle after treatment with liquid nitrogen before lysis). While there are multiple methods for DNA extraction, the most convenient ones in terms of duration, cost, and safety (less use of hazardous chemicals) are silica-based (**Figure 18**). After lysis, DNA binds to a silica membrane

Figure 17. Options for sterilizing forceps/blades after handling each sample. Left: most common setup using a gas burner and ethanol. Right: chemical sterilization using ELIMINase.

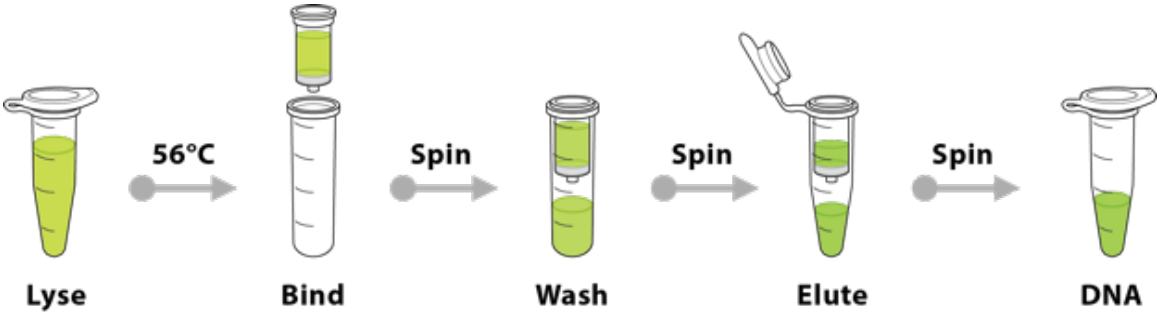


Figure 18. Schematic representation of DNA extraction based on silica.

in the presence of salts and under specific pH conditions. A series of wash steps followed by centrifugation removes proteins, other cellular macromolecules, and salts, allowing pure DNA to be eluted from the membrane into a collection tube for subsequent sequence analysis. One protocol is described in the CCDB (Canadian Centre for DNA Barcoding) Advances Release 1[44].

Negative controls: To assess the purity of reagents used for DNA extraction, especially if they are made in-house or when processing forensic samples, a negative control should always be included. This tube proceeds through all extraction steps but lacks any source of DNA. At the end of the extraction, no DNA should be detected in this negative control. Presence of DNA in a negative control indicates contamination.

DNA Quantification

After DNA extraction, it is good practice to determine the concentration of DNA in the extract. If samples fail to provide DNA through extraction, the step can be repeated using new tissue. If the DNA quantity is low, more DNA can be added by modifying the subsequent amplification PCR reaction. If too much DNA is extracted, it should be diluted prior to PCR.

A spectrophotometer is commonly used to quantify DNA by measuring the absorbance of the DNA sample. DNA is exposed to UV light at 260 nm and a photodetector measures the light passing through it. An extract with a higher DNA concentration absorbs more light than one with little DNA. DNA concentration is calculated based on the absorbance ratio. In addition, the ratio of absorbance at 260 and 280 nm (A260/A280) is used to infer the purity of a DNA sample.

In the absence of a spectrophotometer, DNA extracts can be run with a molecular-weight size marker on an agarose gel (see the section Gel Electrophoresis below) and visualized under UV light.

DNA Preservation

DNA extracts should be stored in a freezer. For short to medium-term storage (on the scale of weeks), household freezers (-20ºC) can be used. For long-term storage, ultra-cold freezers (-80ºC) are preferred. A DNA extract can be aliquoted into multiple tubes to avoid degradation that results from freeze-thaw cycles that occur each time the extract is removed from the freezer to set up a reaction.

In recent years, several countries have established national biobanks. Usually based in natural

44 http://ccdb.ca/resources/

history museums, these repositories provide secure storage of valuable genetic resources for decades. Biobanks such as the National Biodiversity Cryobank of Canada[45] or the U.K.-based CryoArks[46] employ liquid nitrogen storage to hold DNA samples of species of national and international interest (e.g., endangered species).

Polymerase Chain Reaction

The polymerase chain reaction (PCR) makes it possible to amplify a specific DNA region, such as the barcode region, from any genomic DNA extract. This method selectively amplifies the target region, generating the millions of copies needed for analysis through Sanger sequencing. PCR uses temperature cycling to perform polymerase-catalyzed DNA synthesis of the template DNA in the presence of deoxynucleotide triphosphates (dNTPs), primers, and a specific buffer (**Figure 19**). Temperature cycling is performed in a thermocycler (PCR machine), a programmable device that controls the temperature and time of each cycle. After each cycle, the quantity of the amplification target is doubled producing millions of copies of the target DNA region in 25 cycles. Most thermocycling protocols include 30-45 cycles.

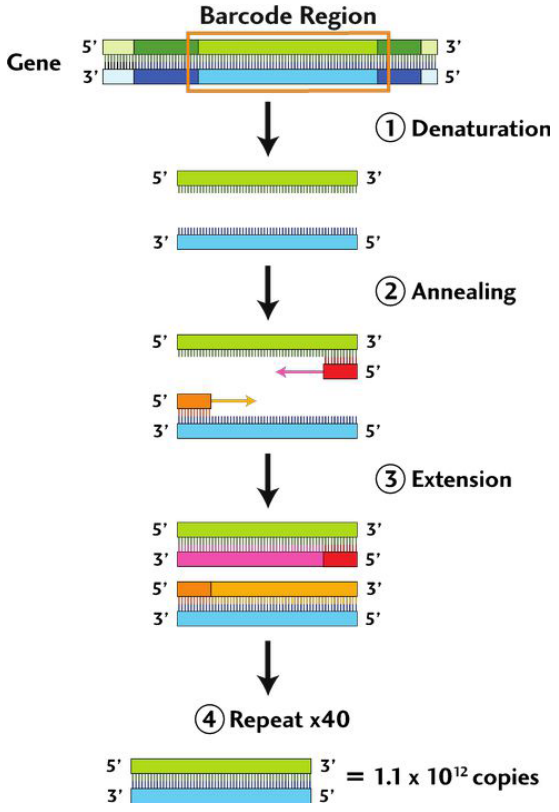


Figure 19. Schematic representation of a PCR reaction for the barcode region. The two DNA strands are coloured in blue and green, while the primers are red and orange.

Each PCR cycle consists of three steps:

1. Denaturation: the hydrogen bonds between the two strands of the DNA helix are broken at high temperatures, usually > 94°C, resulting in two single strands (forward and reverse).
2. Annealing: short synthetic nucleotide sequences (primers), designed to bind to the flanking regions of the target, bind to the single DNA strand at the complementary site under optimized temperature conditions (usually between 45-68°C). The forward (F) primer attaches to the complementary site on the reverse DNA strand while the reverse (R) primer attaches to the complementary site on the forward DNA strand; DNA synthesis is initiated at each of the two primer sites by DNA polymerase.
3. Extension: the polymerase extends the newly synthetized DNA sequence by incorporating the dNTPs present in the PCR mix, creating a new strand with complementarity to the template DNA. The temperature varies between 65°C and 75°C depending on the protocol.

PCR Primers

Primers are short nucleotide sequences (17-30 bp) artificially synthesized to serve as initiation points for the synthesis of the target DNA. Since DNA is double stranded, a pair of primers is used for the simultaneous synthesis of both forward and reverse strands. Primers are **critical** for successful

45 https://nature.ca/en/research-collections/collections/cryobank
46 https://www.cryoarks.org/

amplification of the barcode region. Before starting any barcoding study, literature should be reviewed to identify the most commonly used primers for the taxa under investigation.

Primers can be grouped in a few categories:

- **Species-specific** primers amplify the target region of a single taxon; they are useful in studies that seek to analyze one species.
- **Universal** primers amplify the target region in many taxa; none are truly universal but those with broad application are key for most barcoding studies.
- **Degenerate** primers include multiple nucleotides at certain positions in the primer sequence. Their use increases amplification success in groups with nucleotide variation in the primer binding regions. For example, when a primer lists R at a specific nucleotide site, it means it contains two versions of the primer, one with a C at this site and another with a T at the same site.
- **Primer cocktails** are a mix of different primer pairs amplifying the same target region designed to improve amplification success for a large number of species in a higher taxon (e.g., fish, mammals).
- **Tailed** primers contain extra nucleotides at the 5'-end that are not complementary to the target gene sequence. The 3'-end of the tailed primer anneals to the target region while the non-complementary 5'-end, creates a "tail" of additional nucleotides in the PCR product. The tail (e.g., M13) is added to facilitate sequencing. While the primers used for PCR are typically used for sequencing, the tail is used for sequencing when tailed primers are employed. For instance, LepF1_t1 is a tailed primer where the lepidopteran forward primer (LepF1) has an M13-forward tail (red): 5'-TGTAAAACGACGGCCAGTATTCAACCA ATCATAAAGATATTGG-3'. Tails are used with primer cocktails to allow sequencing in a single reaction. Without tails, each possible primer combination in the cocktail would require its own sequencing reaction.

Primers most commonly used for DNA barcoding of animals, plants, and fungi are listed in BOLD's Primer Database[47].

Gel Electrophoresis

To verify the success of amplification, PCR products are loaded onto an agarose gel immersed in a buffer. An electrical current is applied, and any amplicons plus residual primers will migrate towards the positive electrode. The distance travelled in the gel is determined by the run time and the size of the DNA fragment (large amplicons run slower than short ones). By incorporating a dye which binds with the DNA, the presence and position of the PCR products can be visualized under UV light. A DNA ladder (a mixture of DNA fragments of known size) is often used to estimate the size of the fragments in the gel to verify the target region was amplified.

The Gel electrophoresis workstation should be located on a separate bench and should be considered as a 'contaminated area'. Ethidium bromide, still widely used for the visualization of DNA, is a mutagen which should be treated as a hazardous chemical. Assign separate consumables (pipettes, tips, tube racks, permanent markers, tape, gloves), equipment (electrophoresis set-up, gel imaging system) and coats to this station.

DNA Sequencing

Once PCR amplicons have been generated, the next step involves their sequence characterization. Sanger sequencing was the standard approach from 1980–2010 so construction of the DNA barcode reference library was reliant until recently on this technology. High throughput sequencing (HTS) platforms (Sequel, Sequel II)

47 http://boldsystems.org/index.php/Public_Primer_PrimerSearch

developed by Pacific Biosciences (PacBio) provide a new option as they can generate high fidelity reads for the amplicons employed for DNA barcoding. Furthermore, they can recover reference barcodes from 10,000 specimens per run, enabling a 40-fold cost reduction from Sanger sequencing[48]. As such, the Sequel platforms are currently the most cost-effective technology for constructing a DNA barcode reference library. However, Sanger sequencing remains widely used for processing smaller numbers of specimens, so the balance of this manual considers data generated with this method.

It is often most convenient and economical to send PCR products to a sequencing facility for analysis because DNA sequencers are expensive to purchase and maintain. The sequencing facility will return results as sets of files generated by the DNA sequencer (trace files /chromatograms with Sanger sequencing), each file corresponding to a sample. For reference library building, these files need to be uploaded to BOLD to preserve the raw data. Steps for data upload are detailed in the BOLD Handbook. Users can verify the accuracy of sequence editing by downloading and analyzing these files. HTS platforms do not generate chromatograms which is why only sequence files in text format (fasta) are uploaded.

48 Hebert PDN, Braukmann TWA, Prosser SWJ, et al. 2018. A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19: 219.

Chapter 4.

Sequence Data Management and Analysis

Sequence Editing

Sanger sequencers produce a chromatogram (also known as electropherogram or trace file), from which the DNA sequence is derived. Common file formats for chromatograms are ABI (created by the Applied Biosystems sequencers) and SCF (created by platforms such as Beckman or Li-Cor). These files require specific software to be viewed and analyzed. Free software options include BioEdit[49], Chromas[50] or FinchTV[51], while commercial options include CodonCode[52], Geneious[53] or ChromasPro[54], all of which include extensive analytical tools beyond simple sequence editing.

The barcode region is usually sequenced bi-directionally (forward reaction and reverse reaction) to recover the entire DNA sequence. This is especially important when creating a DNA barcode reference library. Forward and reverse traces are assembled into one contig, representing one sequence, and edited to fix ambiguous bases (if possible). Unidirectional sequencing will recover a slightly shorter DNA fragment which is still very useful when identifying unknown specimens based on a reference library. Regardless of software choice, the steps required to move from raw data (trace file) to a reliable DNA sequence are identical (**Figure 20**). In the case of unidirectional sequencing, no contig can be assembled as there is only one trace file per sample, but all the other editing steps remain the same.

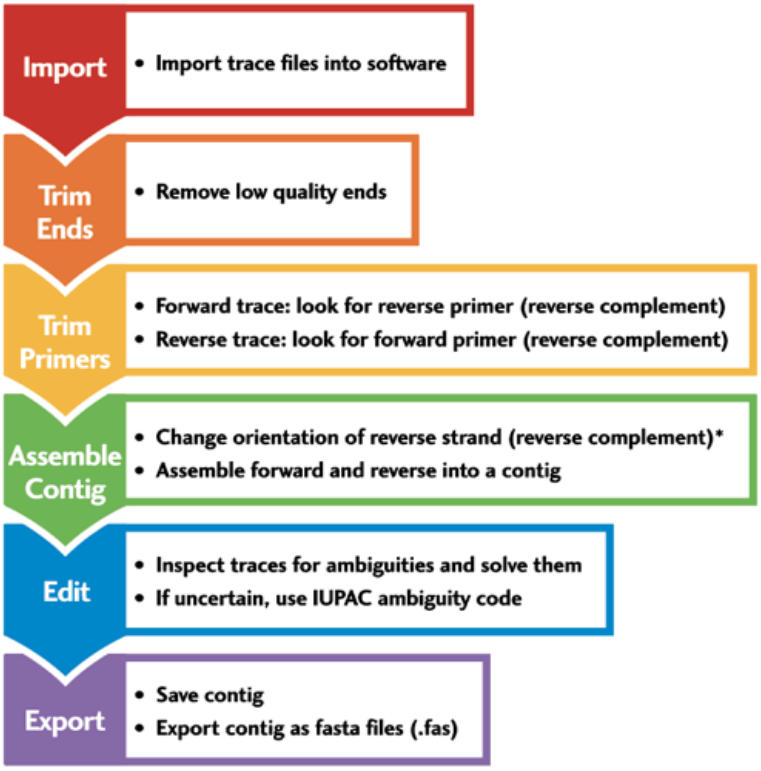


Figure 20. Sequence editing workflow.

FASTA files (.fas) are the most common file type used for storing DNA sequences and they are compatible with most analytical software. FASTA files are identified by the symbol '>' before the sequence name in the first row of a sequence, followed by the string of nucleotides in the second row:

>Unique identifier of the sequence

ACCTGGGCAAATTT

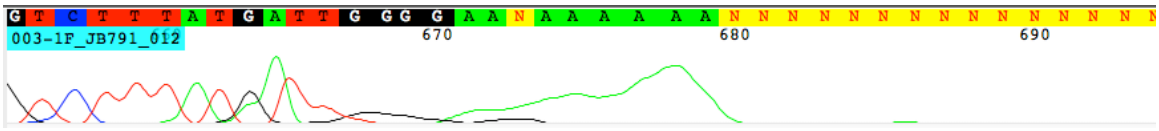
If BOLD is used as a workbench, you should assign either the Sample ID or Process ID as the name of the sequence during editing to facilitate data upload to BOLD.

49 https://bioedit.software.informer.com/
50 http://technelysium.com.au/wp/chromas/
51 https://finchtv.software.informer.com/1.4/
52 https://www.codoncode.com/
53 https://www.geneious.com/
54 http://technelysium.com.au/wp/chromaspro/

Single sequence editing

The thee basic steps for sequence editing are constant, but each software package has specific commands to perform them.

1. Trim ends: the margins of the chromatograms are usually low-quality and need to be removed (usually the software allows for an easy 'highlight and delete' option).

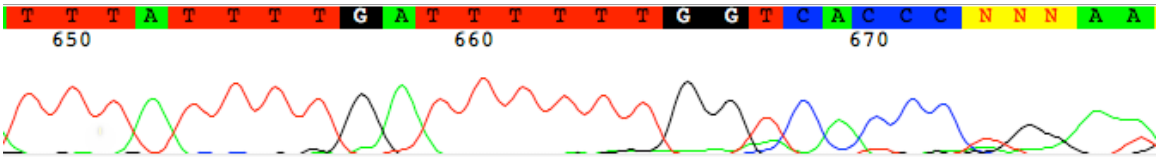


Trim primers: primers are synthetic sequences that bind to the DNA template. Therefore, they may not reflect the true sequence at the site of annealing and should be removed from the final DNA sequence. In many cases, only a fragment of the primer can be reliably observed (see below) due to the low quality of base calls near the end of each chromatogram, but this does not impact the final DNA sequence as primer sequences are always trimmed. In the example below, sequencing was performed with universal invertebrate primers[55] but only the highlighted sections were recovered:

LCO1490 (forward primer): 5'- GGTCAACAAATCATAAAGATATTGG-3'

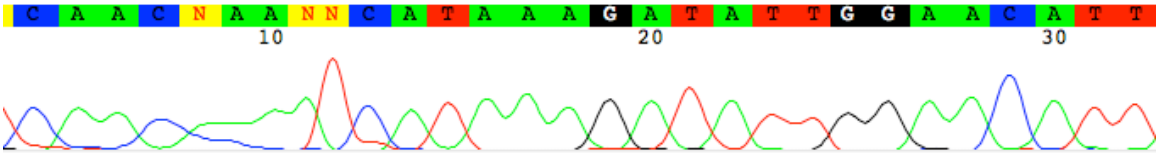
HCO2198 (reverse primer): 5'-TAAACTTCAGGGTGACCAAAAAATCA-3'

Forward trace: go to the end of the forward sequence and look for the reverse primer (as reverse complement).



At position 656, the primer sequence starts: TGATTTTTTGGTCACCC…. which is the reverse complement of HCO2198 (…GGGTGACCAAAAAATCA). Delete primer sequence (starting with position 656 in the example above, until the end of the strand). The remaining sequence should end with TTTATTT.

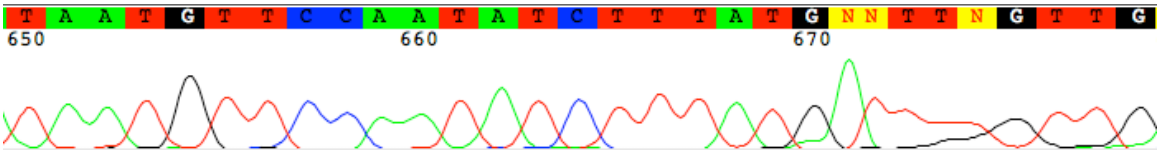
Reverse trace: reverse complement the entire sequence (usually by a click of a button in the software) and look for the forward primer at the beginning of the reverse trace.



At position 26 the primer sequence ends (….CATAAAGATATTGG). Delete primer sequence (in the example above, everything from the beginning of the sequence to position 26). The remaining sequence should start with AACATT.

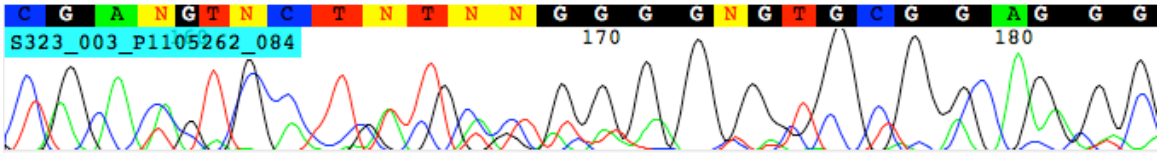
55 Folmer OM, Black WH, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from metazoan invertebrates. *Molecular Marine Biology and Biotechnology* 3: 294–299.

Note: It is also possible to use the reverse trace as it is (not reverse complement and look for the forward primer (as reverse complement) at the end of the sequence (see below). However, the sequence will need to be switched to reverse complement before being assembled into a contig with the forward trace.



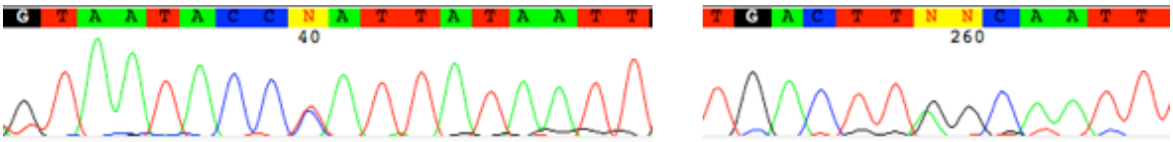
At position 657, the primer sequence begins: CCAATATCTTTATG… which is the reverse complement of LCO 1490 (…CATAAAGATATTGG). In this case, delete everything starting with position 657 until the end of the sequence.

In cases where the low quality of the trace makes it difficult to identify the primer regions, remove the low-quality areas and edit the remainder of the sequence. If the entire trace is low quality, discard it.

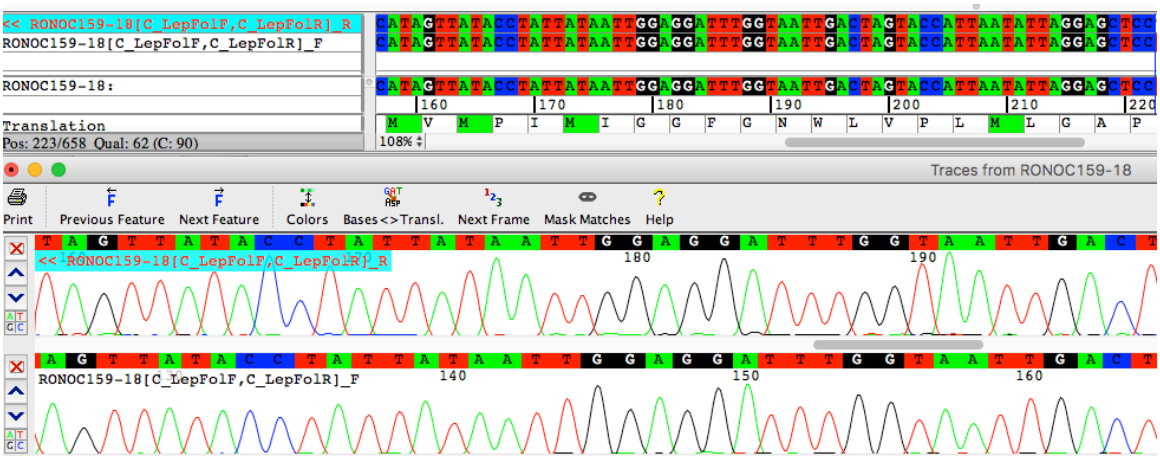


Steps 1 and 2 can also be combined and the necessary fragment trimmed in one step.

2. Sequence assembly and editing: once primers are trimmed, forward and reverse traces are ready for assembly into a single contig (usually by clicking on a button in the software). The contig is ready for manual inspection to verify its quality and to assess ambiguities. If a nucleotide can be 'called' (e.g., a clear peak which erroneously appears as an 'N'), it needs to be corrected. If peaks are overlapping and a decision cannot be made, an IUPAC ambiguity nucleotide code should be used instead. In the example below, the image on the left shows a double peak at position #40 which cannot be resolved while the image on the right shows two ambiguities (positions #259 + 260); the first cannot be resolved (and will stay as 'N') while the second can be called 'G'.



Once the entire contig is verified and edited, the DNA sequence is ready to be exported (as a FASTA file) and used for further analytical steps.



Batch sequence editing

Editing sequences one by one is tedious which is why commercial software supports batch editing. For example, Geneious includes a DNA barcode function that automates all steps from input to contig assembly. In this application, all forward traces are assembled into a master alignment and opened at the same time so their primers can all be trimmed with a click of a button. Similarly, a batch of forward and reverse traces can be merged into individual contigs in one step.

Quality control

When DNA from a broad range of taxa is amplified with universal primers, non-target amplicons (pseudogenes, contaminants) can be recovered. Therefore, it is important to verify the integrity and quality of the sequence to prevent erroneous sequences from being uploaded to BOLD. To accomplish this validation, several steps can be performed with free software. BOLD allows some validation of sequence data upon upload to the workbench (see next section, BOLD Analytics). A popular free software package is MEGA (Molecular Evolutionary Genetic Analysis)[56].

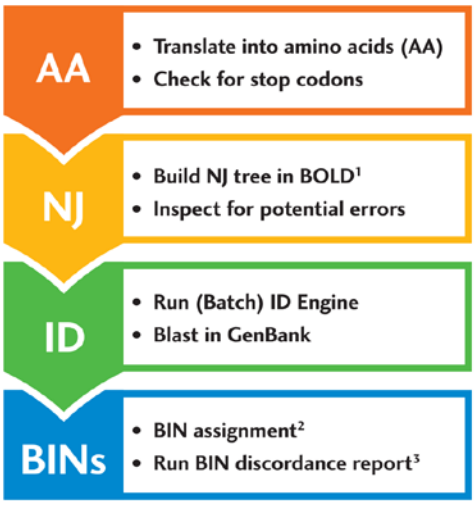


Figure 21. Main steps for the validation of new barcode sequences. AA – amino acids; NJ – neighbour-joining; ID – identification; BIN – Barcode Index Number; 1 – NJ trees require more than 3 sequences; 2 – BIN reference library is updated once a month; 3 – Full BIN discordance report should be performed in BOLD 3.5 (in BOLD 4, the BIN discordance report only considers records within a project not all records in the database).

Data validation is a crucial step because the overall DNA barcoding approach relies on access to reference sequence libraries against which unknown sequences can be compared to identify specimens. BOLD and various external software packages offer tools to help users validate and curate their data (**Figure 21**).

Sequence data processing

All analytical steps can be performed on single specimens (e.g., in forensic cases) or on batches of specimens (standard approach in biodiversity assessments). Batch validation of DNA sequences includes the steps described below:

- Alignment of multiple sequences (in MEGA) to check for indels
- Translation of nucleotides into amino acids (in MEGA) to check for stop codons
- BLAST in GenBank to identify the sequence to aid recognition of potential contamination
- BOLD ID Engine (similar role to BLAST). To allow concomitant multiple queries, a user account is needed. Up to 100 sequences can be queried at a time through the ID Engine. Projects with more than 100 sequences can utilise the Batch ID Engine in BOLD 4 (see the section on BOLD Analytics below)

Translation into amino acids

Within the DNA barcoding pipeline, a DNA sequence (nucleotides) is translated into a protein sequence (amino acids) to check for the presence of stop codons. Their presence indicates either a sequencing error or the inadvertent amplification of a pseudogene because functional proteins do not have internal stop codons. This translation step only applies to protein-coding genes such as the animal barcode region (COI) and the standard plant barcodes (matK and rbcL), but not to ITS.

56 Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* 33: 1870–1874.

A sequence of three nucleotides, referred to as codon, directs the incorporation of a specific amino acid into a protein. The 64 possible triplets of nucleotides (64 codons) code for 20 different amino acids so the code is 'degenerate' because several codons often code for a particular amino acid. A few codons, called *stop codons*, do not code for an amino acid but signal the end of protein synthesis. In the standard genetic code, TAA, TAG and TGA are stop codons. The genetic code for mitochondria varies among different groups of animals so use of the correct genetic code before translating nucleotides into amino acids is crucial.

The translation of a gene is initiated at a start codon and ends with a stop codon. In eukaryotes, there is only one start codon, ATG, encoding the amino acid methionine. The reading frame starts with the letter "A" of the start codon. If it starts from the second letter of the start codon (T), the translation will not be in frame producing incorrect amino acid reads. The barcode region does not start at position 1 of the gene, which means that a minor adjustment is required during translation. For instance, if working with newly edited COI sequences (primers removed) in software such as MEGA, translation will result in many apparent stop codons. To correct it, MEGA users need to insert two dashes before the first nucleotide of each sequence. Translation then produces a sequence of amino acids in correct reading frame. After inspection of the amino acid sequence output, the translation can be reverted to the DNA sequence and the dashes deleted.

The translation step can also be performed for batches of sequences to speed processing.

Sequence alignment

DNA sequences must be aligned using software to ascertain the level of sequence divergence between them. The process of alignment may expose nucleotide **in**sertions and **del**etions (= INDELS) which can reflect sequencing errors, the recovery of a pseudogene, or the deletion/insertion of a codon in the target gene.

If only one specimen is processed, additional sequences can be downloaded from public databases to allow alignment of the new sequence. However, the specimen that generated the new sequence needs at least a family identification to select an appropropriate reference sequence.

Free software is available to support sequence alignment. MEGA offers two algorithms for sequence alignment. If many sequences are being aligned, MUSCLE[57] performs faster than ClustalW[58].

Taxonomic Assignment

GenBank

Once sequences are edited, they can be queried against a database of reference sequences. This process reports the closest taxonomic matches (i.e., the new sequences will receive a species assignment if the database has coverage for it). The most widely used identification tool is the Basic Local Alignment Search Tool (BLAST)[59] in GenBank. This web-based tool locates and displays regions of similarity between DNA sequences while also providing quality scores for sequence matches. BLAST offers two options for DNA barcodes: 1) nucleotide BLAST (blastn[60]), the most commonly used option in DNA barcoding as it

57 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32: 1792–1797.

58 Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22: 4673–4680.

59 Altschul SF, Gish W, Miller W, et al. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

60 Zhang Z, Schwartz S, Wagner L, Miller W. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology* 7: 203–214.

compares the submitted nucleotide sequence with other nucleotide sequences in GenBank, and 2) blastx that translates the nucleotide input into all possible reading frames of protein to compare it to an amino acid sequence database (**Figure 22**). A blastn query can be optimized for different levels of expected similarity. Batch BLAST searches are performed by uploading a fasta file with multiple sequences or pasting multiple sequences in the query box. A BLAST function is included in some sequence editing and analytical software packages (e.g., CodonCode, MEGA) where a (single) sequence can be compared to GenBank through a direct link from the software which connects to the query page of GenBank.

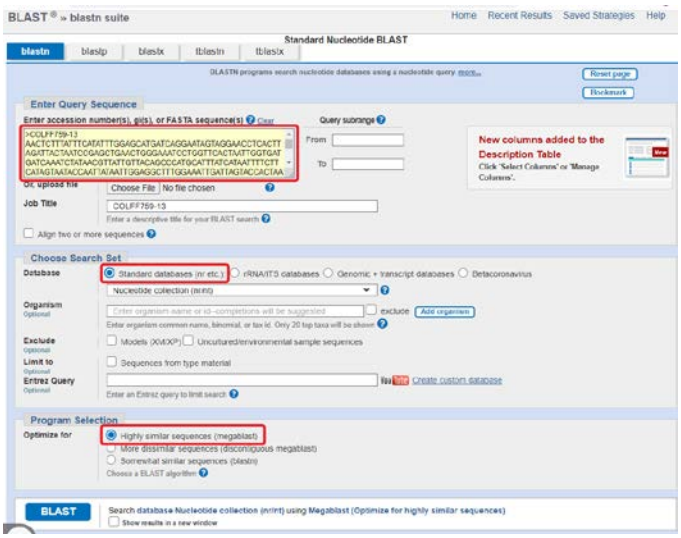


Figure 22. Comparison of a DNA sequence against a database through BLAST. The main steps are circled in red (query box to paste the unknown sequence, choice of database and program selection).

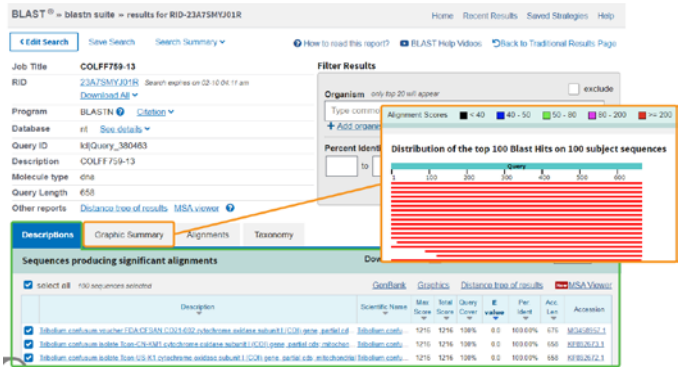


Figure 23. Output of a query with BLAST. There are three main sections with results: 'Graphic Summary' (with colour-coded alignment scores between the query sequence and the closest 100 matches in the database), 'Descriptions' (a list of 100 closest matches with taxonomic identification, query coverage, E-value and % identity) and 'Alignments' (query sequence aligned with each one of the closest 100 matches). Additional reports can be investigated and the results can be downloaded.

A standard BLAST output includes various scores (Max score, Total score, Query cover, E-value, Identity), and the Accession number of the match (**Figure 23**). It also generates reports that include a Search Summary, Taxonomy reports, Distance trees of results, and a MSA viewer (multiple sequence alignment viewer). The results can be downloaded if desired. Each page of results includes a link to a YouTube video with instructions on data interpretation.

BOLD

BOLD offers a sophisticated query tool called the BOLD Identification Engine[61] (ID Engine), allowing single or batch queries of the database. Performing single queries on the ID Engine does not require a user account while batch queries do require one.

As opposed to BLAST, where any DNA sequence can be compared against GenBank, BOLD has been built specifically as a platform for DNA barcoding. Therefore, only standard barcode marker sequences (COI in animals, matK and rbcL in plants, ITS in fungi) can be queried. in the case of animal sequences, BOLD rapidly aligns the query sequence to a global alignment through a Hidden Markov Model (HMM) profile of the COI protein. This is followed by a linear search of the reference library. The user selects from one of four reference databases (**Figure 24**):

- All Barcode Records
- Species Level Barcode Records
- Public Barcode Records
- Full Length Barcode Records

61 http://v4.boldsystems.org/index.php/IDS_OpenIdEngine

The most common choices are either the entire database (includes records only identified to a higher taxonomic level) to see the closest match to a query sequence or the database containing only records with species names (when the goal is to find the closest species name to the query). The sequence is pasted into the query box (**Figure 24**) and the results from this query are displayed in a new window (**Figure 25**). Fungal (ITS) and Plant (matK and rbcL) identifications employ the BLAST algorithm instead of the standard BOLD identification engine.

Reference Library Management Using the BOLD Workbench

The Barcode of Life Data System (BOLD) is an online workbench and database that supports the assembly and use of DNA barcode data. It is both a collaborative hub for the scientific community and a public resource. Once published, barcode data become accessible to anyone.

BOLD is organized in projects which hold specimen records, each composed of two pages which aggregate key information: 1) **Specimen Data** – related to the physical specimen (designated by a unique **Sample ID** given by the user), and 2) **Sequence Data** – associated molecular data (designated by a **Process ID** which is automatically assigned by BOLD on upload). Data submission consists of four parts: specimen data (validated by the BOLD data managers), images, traces, and DNA sequences directly submitted by users with no inspection by BOLD data managers, but with automatic quality checks providing the user with instantaneous feedback on sequence quality and potential errors.

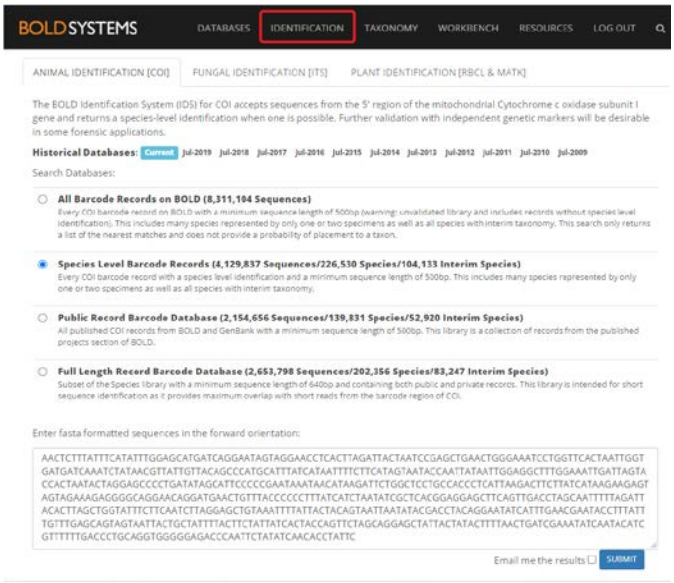


Figure 24. BOLD ID Engine interface showing a single sequence being queried against the Species Level Barcode Records Database. The tool can be accessed directly from the BOLD home page through the top links ("Identification").

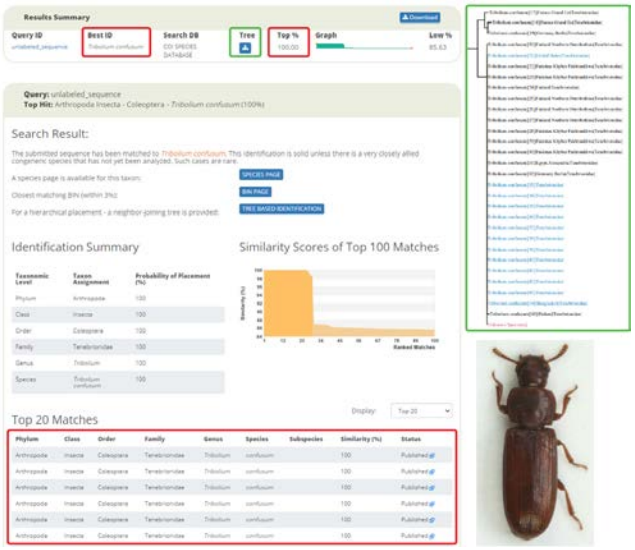


Figure 25. Output page of results from the BOLD ID Engine. The query sequence had very close matches to multiple specimens all assigned to the same species so the identification is considered robust. The results are visualized on a tree (query sequence in red, sequences mined from GenBank in blue) which can be downloaded.

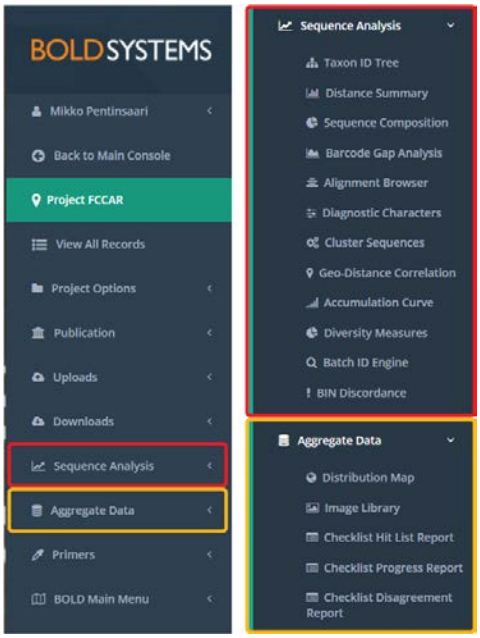


Figure 26. Most of the tools available for data validation and analysis in a BOLD project are situated on the left-side console. By opening the 'Sequence Analysis' and 'Aggregate Data' menus, a suite of tools will be displayed and can be selected with a mouse-click.

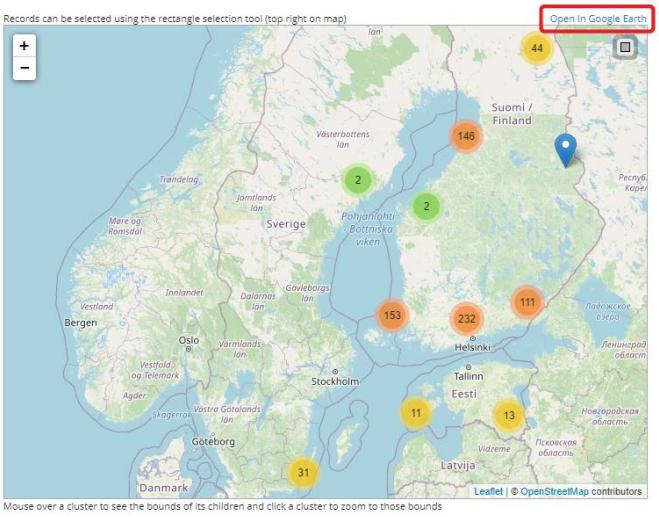


Figure 27. Distribution map for BOLD records from a project targeting North European insects (the red rectangle shows the link to Google Earth).

Data validation

Once a project is populated with records, data can be validated in BOLD (**Figure 26**).

1. After ***specimen data*** upload, a 'Distribution Map' can be built to check the accuracy of the GPS coordinates provided for the specimens in a project (**Figure 27**). The resulting map can be opened in Google Earth[62] for a more detailed view of localities.

- If errors are detected (e.g., incorrect GPS coordinates place specimens in the wrong location), they can be corrected manually (through the Specimen Page of the respective records).
- If there are numerous errors (e.g., all GPS points are incorrect), a batch update can be submitted to BOLD through the specimen data submission interface (see BOLD Handbook for details on data updates).

2. After ***image*** upload, an 'Image Library' can be built to verify that no errors occurred during image submission (**Figure 28**). If an incorrect image has been uploaded, the user must contact BOLD support (support@boldsystems.org) to request its deletion. Once the image has been removed, the correct image can be uploaded.

3. After ***trace*** upload, errors such as a mix-up of traces between records can be resolved by contacting BOLD support to request removal of affected traces. Once the erroneous trace(s) have been removed, the correct ones can be uploaded.

4. Upon ***sequence*** upload, BOLD flags records with stop codons or if the sequence matches any common contaminants (bacteria, human, mouse, pig, etc.; **Figure 29**). Such records are automatically excluded from the database that supports BOLD ID. It is important to note that the genetic code is chosen automatically by BOLD based on the higher-level taxonomy provided with the record.

62 https://www.google.com/earth/

Figure 28. Example of a image library generated by BOLD used to verify taxonomic assignment and to aid the detection of possible mix-ups of images during upload.

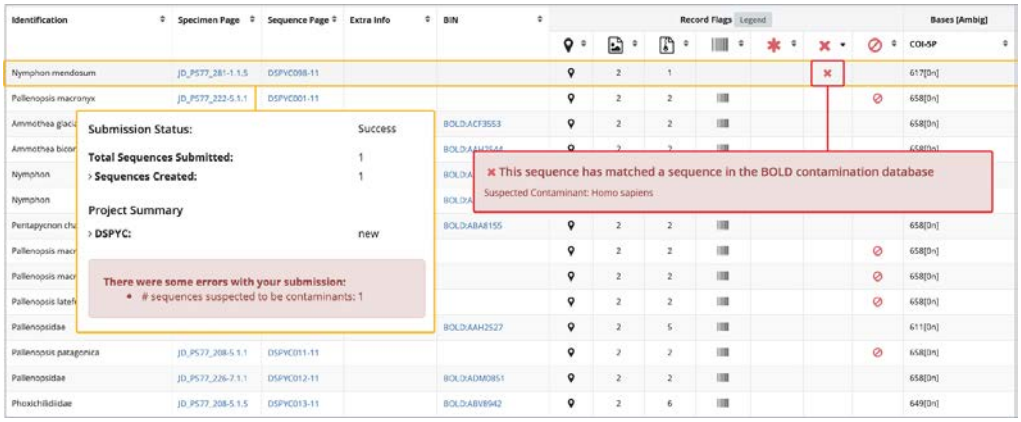


Figure 29. Built-in tools for data validation on BOLD. Contaminants are flagged upon sequence upload to BOLD. Similarly, sequences with stop codons are immediately detected, flagged, and excluded from the database for ID Engine.

5. After sequence upload, the **Batch ID Engine** can be used to run all sequences within a project against one of two databases: all barcode sequences on BOLD, or only barcodes with species designation (**Figure 30**). When datasets are large, it is best to request that BOLD email the results to you. The email will provide a spreadsheet with the 100 closest matches for each sequence queried. If the ID for one or more specimens is clearly incorrect, the following actions should be taken:

- If the conflict reflects an error in data entry for a specimen, update taxonomy manually through the Specimen Page, or if many records are involved as a batch update submitted through the specimen data submission interface (see BOLD Handbook for details).
- If the conflict reflects contamination during analysis, contact BOLD support to request that the record be flagged.

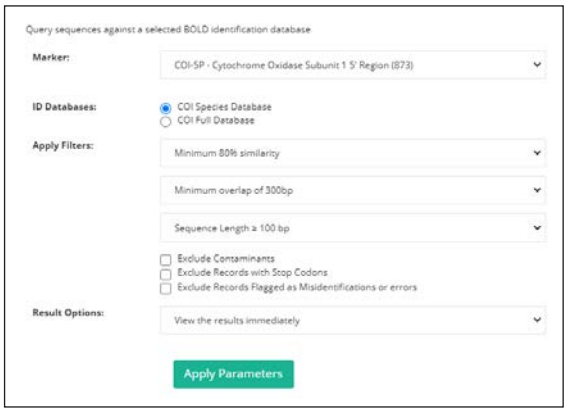


Figure 30. Batch ID Engine allows every record in a project to be compared with the full/species database. Several filters are available. When large numbers of specimens need an identification, have the results emailed to you.

Data analysis

The following analytical tools can be used for data validation and analysis.

1. **Neighbor-joining (NJ) trees** can be built using the Sequence Analysis console by selecting **Taxon ID Tree**. Various parameters (alignment type, genetic distance model) and optional branch labels can be selected (**Figure 31**). The NJ tree allows rapid assessment of the number of barcode clusters and helps to highlight cases that require additional investigation (e.g., occurrence of multiple species names in one cluster or species splitting into multiple clusters). Trees can be downloaded as PDF, Newick, or Postscript files.

2. **Barcode Gap Analysis** provides an overview of the interspecific divergence values among all species in a project, highlighting cases which require additional investigation (intraspecific distances higher than distances to the nearest neighbor species). Users can select varied parameters for analysis, but must designate an alignment method. The results window shows distance values (maximum intraspecific and minimum interspecific) displayed as various scatterplots (**Figure 32**) and summarized in a table format.

3. **Distance analysis.** The divergence values among the DNA sequences in a project can be

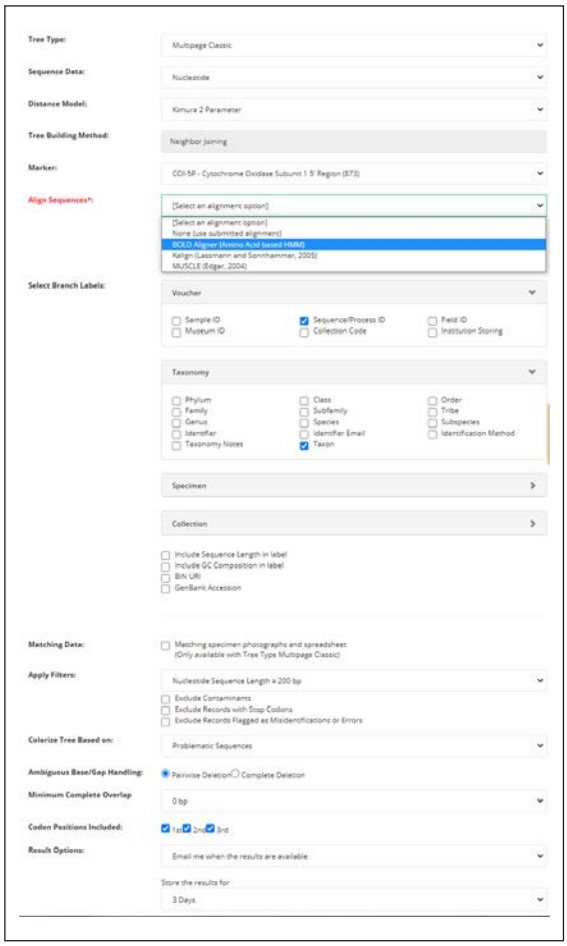


Figure 31. Taxon ID tree settings: only the alignment option is a mandatory field. For the other fields, BOLD will use default settings (process ID and lowest taxonomic level available) unless other options are selected by the user. The tree can be accompanied by a matching image library and a spreadsheet with specimen details. If the data set is large (>1000), have the results emailed.

Note: Although NJ trees are valuable for data validation and barcode visualization, they are not the best approach for infering phylogenetic relationships among taxa.

obtained by selecting **Distance Summary** on the Sequence Analysis console. Various parameters can be adjusted by users with the designation of an alignment method as the only mandatory field. The results of this analysis show the range of distance values (as %) within and between species as well as histograms for these values (**Figure 33**). High intraspecific values (>3%) often indicate potential cryptic species or misidentifications. In the cases of potential cryptic species, additional molecular markers or morphological,

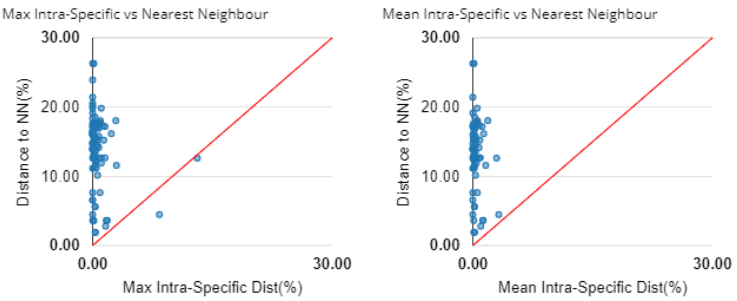


Figure 32. Results from the barcode gap analysis of all records in a project. Scatterplots are built for various analyses; above, the intraspecific distance (maximum or mean) is plotted against the distance to the nearest neighbor species. Dots below the red line indicate species which need further investigation (nearest neighbor is closer than the maximum intraspecific value). Details (distance values as %) for each species and its closest neighbour are provided in a table format and can be downloaded as a spreadsheet.

Figure 33. Results of the distance analysis between DNA sequences in a particular project on BOLD. Divergence values within species and between congeneric species are displayed in both tabular format and as histograms. All values of pairwise comparisons can be downloaded in a spreadsheet.

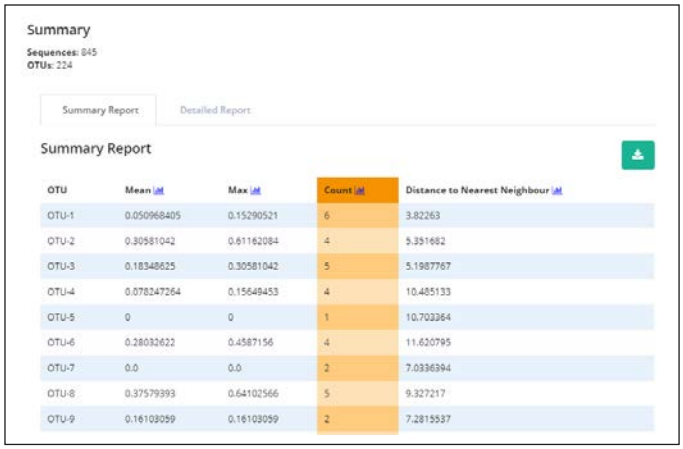
behavioural, or ecological studies are needed to clarify the taxonomic status of the divergent clusters in a species.

4. The **Barcode Index Number** (BIN)[63] system is a valuable tool for evaluating the number of species represented in a dataset, especially for groups where taxonomic study has been limited. BIN assignments are made by BOLD using an algorithm which analyzes all COI sequences on BOLD on a monthly basis and partitions them into groups termed BINs. Each BIN receives a unique identifier (3 letters followed by 4 numbers (e.g. BOLD:AAA1111). BOLD creates a separate page for each BIN that consolidates information on its members (see example BOLD:AAA9566). Due to the strong concordance between BINs and morphological species in groups with advanced taxonomy, BINs are good proxies for species. Each record in a project displays a link to the BIN page on the project console if certain conditions are met (see below). The BIN algorithm is run on the BOLD database monthly so BINs are not provided immediately upon sequence

63 Ratnasingham S, Hebert PDN. 2013. A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS ONE* 8: e66213.

upload. BIN designations are only available for animal COI sequences. Sequences <500 bp are not assigned to a BIN unless they closely match sequences in an existing BIN.

The Cluster Sequences tool will group all DNA sequences from a project into Operational Taxonomic Units (OTUs; **Figure 34**). Although the BIN algorithm is employed, the resultant OTUs are temporary units which do not receive persistent pages in BOLD although data can be downloaded as spreadsheet. In addition, OTUs are project-based while BINs consider the entire BOLD database.



OTU	Mean	Max	Count	Distance to Nearest Neighbour
OTU-1	0.050968405	0.15290521	6	3.82263
OTU-2	0.30581042	0.61162084	4	5.351682
OTU-3	0.18348625	0.30581042	5	5.1987767
OTU-4	0.078247264	0.15649453	4	10.485133
OTU-5	0	0	1	10.703364
OTU-6	0.28032622	0.4587156	4	11.620795
OTU-7	0.0	0.0	2	7.0336394
OTU-8	0.37579393	0.64102566	5	9.327217
OTU-9	0.16103059	0.16103059	2	7.2815537

Figure 34. Results of OTU clustering in a BOLD project. Mean and maximum values for OTUs as well as the distance to the nearest OTU are displayed and can be downloaded (green button). OTUs are numbered but are not persistent.

Figure 35. BIN discordance report for a BOLD project (in BOLD 3.5). For the discordant BINs, the rank of discordance and details on conflicting records are reported. Results can be downloaded by clicking on the green button.

Aside from their role in assigning barcode sequences to provisional species, BINs have an important role in data validation. The **BIN Discordance** Report provides an overview on the (dis)agreement of data within a project to other data on BOLD (**Figure 35**).

Data Publication

Every barcode record on BOLD resides in just one project. However, the individual data records can be partitioned (even a subset of the data from one project) or merged (data from different projects) in **datasets** (**Figure 36**). Datasets enable usage of the same data in multiple studies. Datasets come with the same options as projects and can be analyzed with the tools discussed in the preceding section. Once data is ready for publication, a user can release the entire dataset to the BOLD Public Database as well as to GenBank (see **Figure 36**). At the same time, a DOI (Digital Object Identifier)[64] can be requested and included in the publication (the DOI link allows rapid access to the dataset directly from the publication).

Note: The data owner (i.e. the person who manages the project with the data record) is responsible for public release of the data in BOLD upon publication (ideally even earlier) and for ensuring the data has been submitted to GenBank. As well, citation details for any publication associated with the dataset should be uploaded to BOLD once it is published.

Figure 37 summarizes the analytical steps required for data validation, analysis, and publication in BOLD. For more details on data validation, analysis, and management, see the BOLD handbook (http://www.boldsystems.org/index.php/resources/handbook?chapter=7_validation.html).

64 https://www.doi.org/

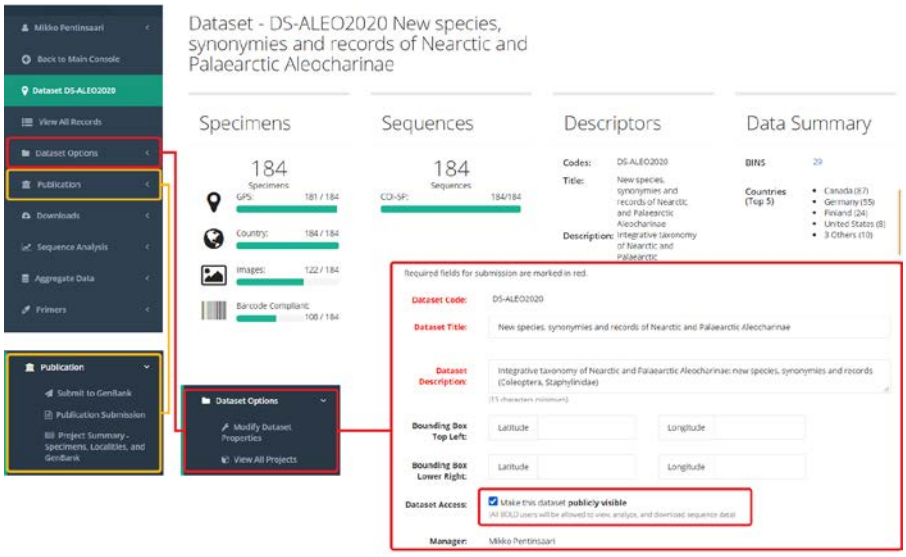


Figure 36. Workflow for publishing datasets. Data should be submitted to GenBank through Publication → Submit to GenBank (yellow rectangles). The dataset should also be publicly released in BOLD through Dataset Options → Modify Dataset Properties → Make this dataset publicly visible (check the box; red rectangles) → Save. A new window appears where a request for DOI can be made.

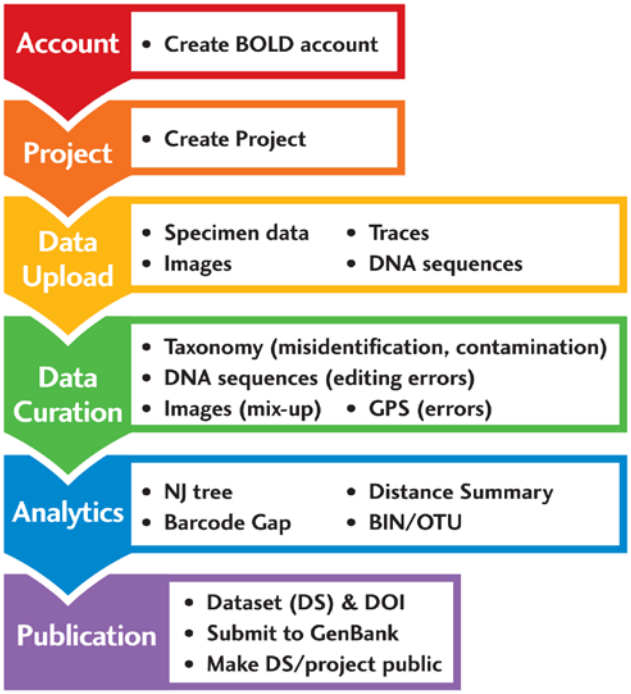


Figure 37. Main workflow and steps required for data validation, analysis, and publication in BOLD. NJ – neighbor-joining; BIN – Barcode Index Number; OTU – Operational Taxonomic Unit. Additional analytical tools are available on BOLD. Once data is ready to be published, it must be submitted to GenBank and also publicly released in BOLD. BOLD supports the assignment of a DOI for each Dataset.

Chapter 5.

Application of DNA Barcoding to Biosurveillance

The instructions for DNA barcoding outlined in the previous chapters can be employed for diverse applications. Since environmental change is currently restructuring ecosystems on a planetary scale, the International Barcode of Life (iBOL) Consortium has prioritized global biosurveillance and activated the BIOSCAN program[65] in June 2019 to implement it. This chapter briefly describes several components of biosurveillance that can be addressed by coupling DNA barcoding with the capabilities of current high-throughput sequencing platforms.

Detecting Invasive Alien Species, Pests and Vectors

Invasive alien species (IAS) are species introduced, intentionally or accidentally, to areas outside their natural distribution often with risks to native species and ecosystems. In addition to the economic and social damage cause by IAS, they are one the main drivers of biodiversity loss[66]. Several online databases maintain verified inventories of IAS, including the Global Register of Invasive and Introduced Species (GRIIS)[67] and the global list of the 100 worst IAS[68]. The top 100 IAS include species such as the Nile perch (*Lates niloticus*), the water hyacinth (*Eichhornia crassipes*), and the zebra mussel (*Dreissena polymorpha*).

Pest species, both native and introduced, causing indirect damage through their role in the transmission of diseases or direct damage by attacking plants and animals. Organisms causing ecological and economic impact to crops, forestry, and gardens have greatly increased their distribution and impact in recent decades due to the globalization of trade. The International Plant Protection Convention (IPPC) focuses on protecting the plant resources from pest species while ensuring safe trade[69]. IPPC sets global standards for plant health and, as such, it has adopted International Standards for Phytosanitary Measures (ISPMs) including detailed diagnostic protocols for regulated pests (ISPM 27)[70]. Some examples of important pests are the desert locust (*Schistocerca gregaria)*, Mountain pine beetle (*Dendroctonus ponderosae*), and *Phytophthora infestans (fungus causing potato blight).*

Vectors are organisms which transmit diseases to plants, animals, and humans that are mediated by the pathogens which they carry. Some well-known vectors are mosquitoes (which are responsible for the transmission of many important human diseases, e.g., Zika, malaria, dengue, chikungunya; **Figure 38**), ticks (carrying bacteria responsible for Lyme disease, and viruses causing encephalitis), and biting midges (transmitting the bluetongue disease to sheep, cattle, and wildlife). The World Animal Health Organization (OIE)[71] oversees animal health and maintains a list of diseases (117 in 2020) of importance for international trade. The diagnostic tests for diseases now routinely include molecular methods to identify pathogens and also use it to identify vectors helping to prevent outbreaks by enabling rapid vector control measures.

The economic impact of IAS, pests, and vectors, through both their direct damage and the cost of the measures to control or eradicate them, likely approaches a trillion dollars annually. For instance, the cost of IAS totals $120 billion per

65 https://ibol.org/programs/bioscan/
66 https://ipbes.net/global-assessment
67 http://www.griis.org
68 Lowe S, Browne M, Boudjelas S, De Poorter M. 2000. 100 of the World's Worst Invasive Alien Species. A selection from the Global Invasive Species Database. The Invasive Species Specialist Group (ISSG), World Conservation Union (IUCN), 12pp.
69 https://www.ippc.int/en/
70 FAO (Food and Agriculture Organization). 2006. ISPM no. 27: diagnostic protocols for regulated pests. International Standards for Phytosanitary Measures 1 to 29 (2007 edition), Secretariat of the International Plant Protection Convention, Rome, pp 341–352.
71 https://www.oie.int/

Figure 38. A bulk trap sample of mosquitoes and other insects collected in Ghana as part of the Target Malaria[72] project that will DNA barcode bloodmeals of biting insects known to be vectors of malaria and other diseases.

year for the United States[73] and $10 billion per year for Australia[74]. Rapid and accurate species identification is vital to avoid the establishment of IAS in new areas, to ensure safe trade of goods and commodities, and to help prevent outbreaks of diseases. Each nation has a list of priority species in this regard and measures in place to prevent, control or eradicate IAS, pests, and vectors. Despite their limitations, diagnostic protocols remain heavily reliant on morphological approaches for plants, animals, and fungi. By contrast, the detection and identification of bacteria and viruses has already transitioned to rapid and objective diagnostic tests based on serological and molecular methods.

Soon after DNA barcoding was proposed, it was found to be highly effective in identifying IAS and pests[75] while also meeting or exceeding the standards for diagnostic protocols under ISPM 27[76], and to similarly aid the identification of vectors[77]. Regulatory agencies have been slow to accept, standardize, and implement DNA barcode protocols, but 23 of the 29 diagnostic protocols adopted by IPPC through 2020 include some type of molecular identification (restriction fragment length polymorphism, real-time PCR, reverse transcription PCR, or targeted amplicon sequencing)[78]. In addition, one of the IPPC regional nodes, the European and Mediterranean Plant Protection Organization (EPPO) has adopted a standard (EPPO Standard PM 7/129 (1))[79] with detailed protocols for the identification of arthropods, bacteria, fungi, plants, nematodes and phytoplasmas. It recommends the use of three online databases: GenBank, BOLD and Q-Bank; the last is a small database[80] (1,820 species in 2020) specifically developed to hold

72 https://targetmalaria.org/
73 Pimentel D, Zuniga R, Morrison D. 2005. Update on the environmental and economic costs associated with alien invasive species in the United States. *Ecological Economics* 52: 273–288.
74 Hoffmann BD, Broadhurst LM. 2016. The economic cost of managing invasive species in Australia. *NeoBiota* 31: 1–18.
75 Armstrong KF, Ball SL. 2005. DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 360: 1813–1823.
76 Floyd R, Lima J, deWaard JR, et al. 2010. Common goals: incorporating DNA barcoding into international protocols for identification of arthropod pests. *Biological Invasions* 12: 2947–2954.
77 Besansky NJ, Severson DW, Ferdig MT. 2003. DNA barcoding of parasites and invertebrate disease vectors: what you don't know can hurt you. *Trends in Parasitology* 19: 545–546.
78 https://www.ippc.int/en/core-activities/standards-setting/ispms/
79 EPPO (European and Mediterranean Plant Protection Organization). 2016. EPPO Standard PM 7/129 (1) DNA barcoding as an identification tool for a number of regulated pests. *EPPO Bulletin* 46: 501–537.
80 https://qbank.eppo.int/

Figure 39. DNA barcoding is routinely performed to ascertain if seized specimens or products derive from CITES-listed species.

multi-marker data for quarantine pests of importance to the European Union, developed through the Quarantine Barcode of Life project[81]. Further uptake by regulatory agencies is certain as laboratory infrastructure is developed, as training is provided, and as the DNA barcode reference library needed to support identification is extended through iBOL.

Detecting Endangered Species

The recent IPBES report[82] noted that 75% of land and 66% of ocean areas have been "significantly altered" by human activities, largely as a result of efforts to feed expanding human populations. Because of these habitat modifications, a million species of plants and animals are at risk of extinction. The International Union for Conservation of Nature (IUCN) maintains the Red List of Threatened Species[83], the authoritative registry for species of animals, fungi, and plants at risk of extinction. To ensure international trade does not further diminish populations of these species, the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)[84] has produced checklists that restrict trade in more than 36,000 species of animals and plants. Despite this Convention, illicit trade in wildlife ($10 billion per annum) and timber ($7 billion) has not been halted. DNA barcoding can help to curb this illegal trade by providing a rapid, objective method to identify animals[85] and plants[86], even when only their parts or derivatives (e.g., meat, skin, and organs; powder, seeds, and wood products) can be analyzed. This identification capacity is critical to determine if intercepted material belongs to a CITES-listed species, providing the evidence needed to prosecute traffickers (**Figure 39**).

81 https://www.qbol.org/en/qbol.htm
82 IPBES (Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services). 2019. *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES secretariat, Bonn, Germany, 56 pp.
83 https://www.iucnredlist.org/
84 https://www.cites.org/
85 Steinke D, Bernard AM, Horn RL, et al. 2017. DNA analysis of traded shark fins and mobulid gill plates reveals a high proportion of species of conservation concern. *Scientific Reports* 7: 9505.
86 Lahaye R, van der Bank M, Bogarin D, et al. 2008. DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences* 105: 2923–2928.

Figure 40. A Malaise trap deployed on the tundra in Nunavut, Canada for biomonitoring arthropods as part of the Arctic BIOSCAN project.

Biomonitoring

The impending mass extinction has created an urgent need to inventory biological communities to allow more refined evaluations of trends in biodiversity. Traditional biomonitoring programs have coupled standardized sampling with morphological identification of organisms. Because of the difficulty in identifying a broad spectrum of species, work has often targeted indicator species whose presence, absence, or change in abundance is presumed to indicate changes in the community. However, this process does not scale to the global level and overlooks most of the 10+ million species of multi-cellular organisms that inhabit our planet. Comprehensive surveys on species richness, abundance, and distributions – at a global scale and repeated continuously – are essential to better understand the drivers of biotic change. The coupling of high-throughput sequencing with DNA barcoding (as well as its extensions: metabarcoding and the analysis of environmental DNA; see Chapter 6) has revealed a path that will enable global biomonitoring. The National Ecological Observatory Network (NEON) in the United States now includes DNA barcoding as a core component of their terrestrial biomonitoring program with a focus on selected insect groups (ground beetles, mosquitoes), as part of its 30-year initiative[87]. Several other large-scale biomonitoring projects employing DNA barcoding and metabarcoding have recently gained support including LIFEPLAN[88] (Finland), BioAlfa[89] (Costa Rica), ARISE[90] (Netherlands), and Arctic BIOSCAN[91] (Canada; **Figure 40**). A framework for connecting the 'genomic observatory' sites within these projects has also been developed recently[92] to ensure global integration and to maximize synthetic outputs.

87 Gibson CM, Kao RH, Blevins KK, Travers PD. 2012. Integrative taxonomy for continental-scale terrestrial insect observations. *PLoS ONE* 7: e37528.
88 https://www.helsinki.fi/en/projects/lifeplan
89 https://www.gdfcf.org/bioalfa-bioliteracy-costa-rica
90 https://ibol.org/news/bioscan-partners-awarded-18-million-euros-to-identify-the-full-breadth-of-biodiversity-in-the-netherlands/
91 https://arcticbioscan.ca/
92 Arribas P, Andújar C, Bidartondo MI, et al. 2021. Connecting high-throughput biodiversity inventories - opportunities for a site-based genomic framework for global integration and synthesis. *Molecular Ecology*. In press.

Chapter 6.

Future Directions

The 2030 guide to DNA barcoding will certainly describe approaches and methods that have not yet been proposed as technological advances open new opportunities. This chapter briefly considers some emergent themes that will gain further importance over the coming decade. The rise of DNA metabarcoding, environmental DNA, on-site DNA barcoding, citizen science, and technological advances will all extend DNA barcoding in important ways.

DNA Metabarcoding

DNA barcoding based on Sanger sequencing energized studies in systematics, molecular ecology, taxonomy, and other fields of biodiversity science. Since 2003, DNA barcode sequences for hundreds of thousands of described species have been added to the rapidly growing global reference library, and tens of thousands of new species have been discovered. However, DNA barcoding is based on processing individual specimens, prepared and databased one at a time to maintain the link between voucher and DNA barcode record. Scaling-up to 96-well and 384-well microplates increased throughput but was too expensive to allow the analysis of bulk samples (e.g., from Malaise traps, plankton tows, etc.). New sequencers combined with rapid advances in bioinformatics now allow the inventory of biodiversity at previously impossible scales. Metabarcoding (**Figure 41**) couples high-throughput sequencing and DNA barcoding to identify multiple species from mass collections or from complex samples of environmental DNA[93]. This method is gaining wide use and has the potential to enable global biosurveillance. Metabarcoding is reviewed in detail in a recent CBD SBSTTA document[94].

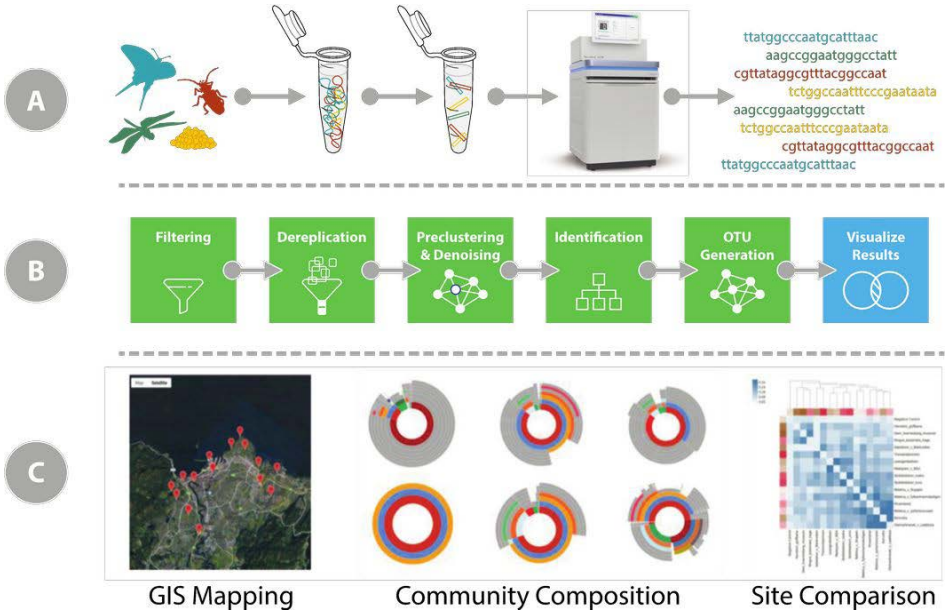


Figure 41. Three main stages in the DNA metabarcoding workflow: (A) molecular analysis involves DNA extraction from a bulk sample sample, PCR amplification of the barcode region, tagging with unique molecular identifiers, followed by pooling and high-throughput sequencing; (B) bioinformatic analysis (using a platform such as mBRAVE) may involve filtering, dereplication, denoising, identification, and generation of operational taxonomic units; and (C) data visualization may include GIS mapping of species composition, measures of community composition, and quantification of change through time at single sites or among sites.

93 Cristescu ME. 2014. From barcoding single individuals to metabarcoding biological communities. *Trends in Ecology and Evolution* 29: 566–571.

94 CBD (Convention on Biological Diversity). 2019. UNEP/CBD/SBSTTA/23/INF/07. Available at: https://www.cbd.int/doc/c/9923/6136/fcaade885666c2cf84d74076/sbstta-23-inf-07-en.pdf

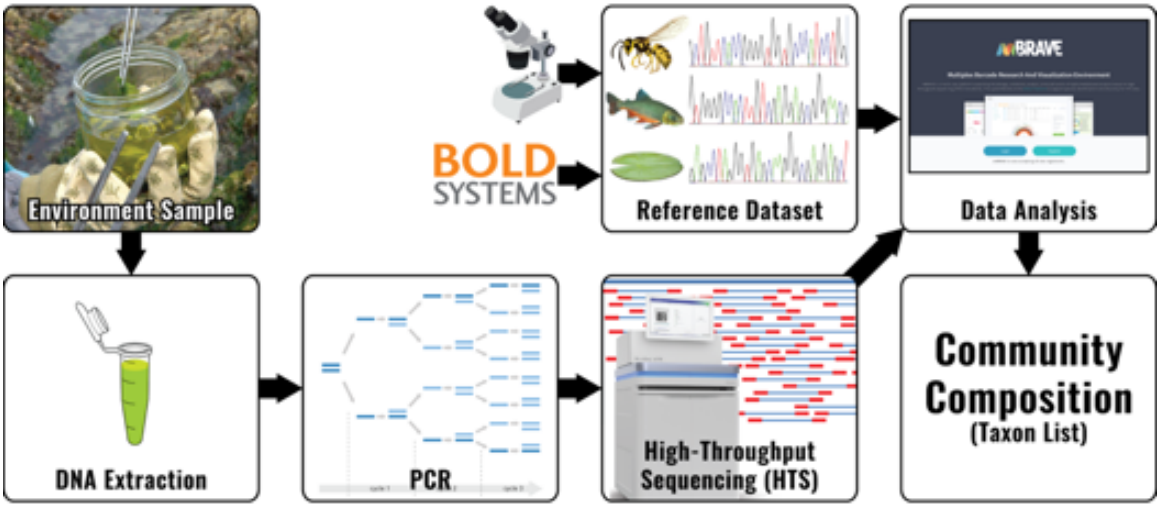


Figure 42. Simplified representation of an eDNA metabarcoding workflow of water samples with data analysis performed on BOLD and mBRAVE[95,96].

Environmental DNA

Environmental DNA (eDNA) is collected from soil, water, air, or snow instead of sampling organisms directly. This approach has the advantage of non-invasive sampling and species identification in near real-time[97]. For instance, the presence of introduced or endangered aquatic species can be ascertained by filtering water and processing DNA captured by the filter. Environmental DNA studies can search for a particular species or can assess community composition. The protocol can allow species detection through real-time PCR, DNA barcoding or metabarcoding (**Figure 42**). As with metabarcoding bulk samples, the use of eDNA has surged in recent years[98].

On-site DNA Barcoding

Recent technological advances have enabled on-site processing and DNA barcoding of samples. DNA sequencers have been miniaturized so they now allow sample processing at the site of collection. Species identification based on real-time PCR has been employed in field settings to detect pests[99] and to inventory biodiversity hotspots[100] by using portable PCR cyclers, but inexpensive portable sequencers (e.g., the Oxford Nanopore MinION[101]; **Figure 43**) with all consumables necessary for sample analysis now fit in a small suitcase (e.g., Bento lab[102]). This miniaturization opens possibilities for point-of-contact species identification, an important development for quarantine pests that need to be diagnosed at port-of-entry and for direct

95 http://www.mbrave.net/
96 Ratnasingham S. 2019. mBRAVE: The Multiplex Barcode Research And Visualization Environment. *Biodiversity Information Science and Standards* 3: e37986.
97 Cristescu ME, Hebert PDN. 2018. Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics* 49: 209–230.
98 Taberlet P, Bonin A, Zinger L, Coissac E. 2018. *Environmental DNA for Biodiversity Research and Monitoring*. Oxford University Press, Oxford.
99 Naaum AM, Foottit RG, Maw HEL, Hanner R. 2014. Real-time PCR for identification of the soybean aphid, *Aphis glycines* Matsumura. *Journal of Applied Entomology* 138: 485–489.
100 Pomerantz A, Peñafiel N, Arteaga A, et al. 2018. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience* 7: giy033.
101 https://nanoporetech.com/products/minion
102 https://www.bento.bio/

Figure 43. The portable DNA sequencer by Oxford Nanopore, the MinION, is a pocket-sized device that permits simple and rapid sequencing on a benchtop, at the sample source, or almost anywhere.

seizures of endangered species. On-site DNA barcoding may also aid expansion of the DNA barcode reference library in biodiversity-rich countries where laboratory infrastructure is limited, and the export of biological specimens is prohibited.

Citizen Science

Citizen science has become an important source of biodiversity data. Often supported solely by a mobile phone application, citizen scientists are tracking wildfires, diseases, marine debris, stream levels, land-use changes, storm damage, light pollution, and illegal logging. They are also sampling organisms or their environments for barcode analysis. For example, the School Malaise Trap Program engaged 15,000 students at 350 schools across Canada to collect over 80,000 insects for DNA barcoding to inventory biodiversity in their schoolyards[103] (**Figure 44**). A recently initiated project in Costa Rica, BioAlfa[104], will involve citizen scientists to collecting specimens for DNA barcode as part of the effort to barcode every species in the country. The coupling of DNA barcoding and an army of citizen scientists equipped with informatics tools and technologies such as artificial intelligence, remote sensing, and autonomous vehicles will undoubtedly speed the inventory of life on Earth[105].

Technological Advances

Much like the first decade of DNA barcoding[106], the next decade will see remarkable advances in protocols enabled by new tools and equipment. The processing of specimens and bulk samples will increasingly integrate emerging technologies for image acquisition and analysis. Automated

103 Steinke D, Breton V, Berzitis E, Hebert PDN. 2017. The School Malaise Trap Program: Coupling educational outreach with scientific discovery. *PLoS Biology* 15: e2001829.
104 https://news.mongabay.com/2020/04/bold-project-hopes-to-dna-barcode-every-species-in-costa-rica/
105 https://news.mongabay.com/2019/03/combining-artificial-intelligence-and-citizen-science-to-improve-wildlife-surveys/
106 Hebert PDN, Hollingsworth PM, Hajibabaei M. 2016. From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371: 20150321.

Figure 44. A Malaise trap deployed on a schoolyard for the School Malaise Trap Program. For this citizen science project, the Centre for Biodiversity Genomics teamed up with thousands of students and educators across Canada to explore the arthropod diversity found in their schoolyards using DNA barcoding.

digital microscope systems (e.g., Keyence VHX-7000; **Figure 45**) now permit the imaging of hundreds of specimens per hour and their automated export into BOLD to support barcode analysis. These specimen images will soon be analyzed by artificial intelligence systems to automate higher-level taxonomic classifications[107] that will be refined through DNA barcoding.

Advances in liquid handlers are improving throughput (e.g., Beckman Coulter Biomek i7) while the use of acoustics to move reagents will shrink reaction volume lowering costs (e.g., Beckman Coulter Echo 525). The adoption of plasma cleaning devices (e.g., Ion Field Tip Charger) to remove residual DNA will allow reuse of plasticware. High-throughput sequencers will continue to advance, aiding both barcoding and metabarcoding workflows. For example, the Sequel and Sequel II platforms from Pacific BioSciences (**Figure 45**) have greatly reduced the costs for barcode analysis by enabling from 10,000-40,000 specimens to be analyzed in each run[108]. Similar protocols have also been released

Figure 45. Technological advance such as automated digital imaging systems (Keyence VHX-7000) and high-throughput sequencers (PacBio Sequel II) are speeding construction of the global DNA barcode reference library.

107 Ding W, Taylor G. 2016. Automatic moth detection from trap images for pest management. *Computers and Electronics in Agriculture* 123: 17–28.

108 Hebert PDN, Braukmann TWA, Prosser SWJ, et al. 2018. A Sequel to Sanger: amplicon sequencing that scales. *BMC Genomics* 19: 219.

for Oxford Nanopore[109] and Illumina[110]. These platforms will make it possible to extend beyond the usual suite of barcode markers in cases where single-locus barcoding is insufficient (e.g., vascular plants[111]). Genome skimming, where a shallow pass of the genome is sequenced using HTS, holds promise for species identification in such groups[112].

109 Srivathsan A, Baloğlu B, Wang W, et al. 2018. A MinION-based pipeline for fast and cost-effective DNA barcoding. *Molecular Ecology Resources* 18: 1035–1049.

110 de Kerdrel GA, Andersen JC, Kennedy SR, et al. 2020. Rapid and cost-effective generation of single specimen multilocus barcoding data from whole arthropod communities by multiple levels of multiplexing. *Scientific Reports* 10:78.

111 Coissac E, Hollingsworth PM, Lavergne S, Taberlet P. 2016. From barcodes to genomes: extending the concept of DNA barcoding. *Molecular Ecology* 25: 1423–1428.

112 Bohmann, K, Mirarab, S, Bafna, V, Gilbert, MTP. 2020. Beyond DNA barcoding: The unrealized potential of genome skim data in sample identification. *Molecular Ecology* 29: 2521– 2534.

Annex

Annex 1: Glossary of Terms and Definitions

ABS	Access and Benefit-Sharing
Arctic BIOSCAN	Project assembling a DNA barcode library for Arctic species.
ARISE	Biodiversity infrastructure and monitoring program in the Netherlands.
BIN	Barcode Index Number
BioAlfa	Acronym for "BioAlfabetizado", a project gathering DNA barcodes for every species in Costa Rica.
BIOSCAN	Research program led by the International Barcode of Life Consortium.
BLAST	Basic Local Alignment Search Tool
Blastn	Nucleotide BLAST
BOLD	Barcode of Life Data System
CBD	Convention on Biological Diversity
CBG	Centre for Biodiversity Genomics
CITES	Convention on International Trade in Endangered Species
COI	Cytochrome *c* oxidase subunit I
DDBJ	DNA Data Bank of Japan
DNA	Deoxyribonucleic acid
DOI	Digital Object Identifier
eDNA	Environmental DNA
EMBL-EBI	European Bioinformatics Institute
ENA	European Nucleotide Archive
EPPO	European and Mediterranean Plant Protection Organization
FAO	Food and Agriculture Organization
.fas	FASTA files
GenBank	Database of nucleotide sequences
GTI	Global Taxonomy Initiative
GTI-DNA-tech	Global Taxonomy Initiative DNA Technologies training program
GRIIS	Global Register of Invasive and Introduced Species
HMM	Hidden Markov Model
HTS	High Throughput Sequencing

IAS	Invasive alien species
iBOL	International Barcode of Life Consortium
ID Engine	Identification Engine in the Barcode of Life Data System (BOLD)
IPBES	Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services
INSDC	International Nucleotide Sequence Database Collaboration
IPPC	International Plant Protection Convention
ISSG	Invasive Species Specialist Group
ISPM	International Standards for Phytosanitary Measures
ITS	Internal Transcribed Spacer
IUCN	International Union for the Conservation of Nature
LIFEPLAN	A Planetary Inventory of Life research project led by the University of Helsinki.
matK	Maturase K
MEGA	Molecular Evolutionary Genetic Analysis
MSA Viewer	Multiple sequence alignment viewer
NBSAPs	National Biodiversity Strategy and Action Plans
NEON	National Ecological Observatory Network
NCBI	National Centre for Biotechnology Information
NUMTs	Mitochondrial sequences in the nuclear genome
NUPTs	Plastid sequences in the nuclear genome
NJ	Neighbor-joining
OIE	World Organization for Animal Health
OTUs	Operational taxonomic units
PCR	Polymerase chain reaction
rbcL	Ribulose-1,5-bisphosphate carboxylase/oxygenase
RNAlater	Reagent that stabilizes RNA
SPHNC	Society for the Preservation of Natural History Collections
SDGs	Sustainable Development Goals
trnH-psbA	intergenic spacer region in plastid genome of plants

Secretariat of the Convention on Biological Diversity

World Trade Centre
413 St. Jacques Street, Suite 800
Montreal, Quebec, Canada H2Y 1N9

Phone: +1 514 288 2220
Fax: +1 514 288 6588
E-mail: secretariat@cbd.int
Website: www.cbd.int