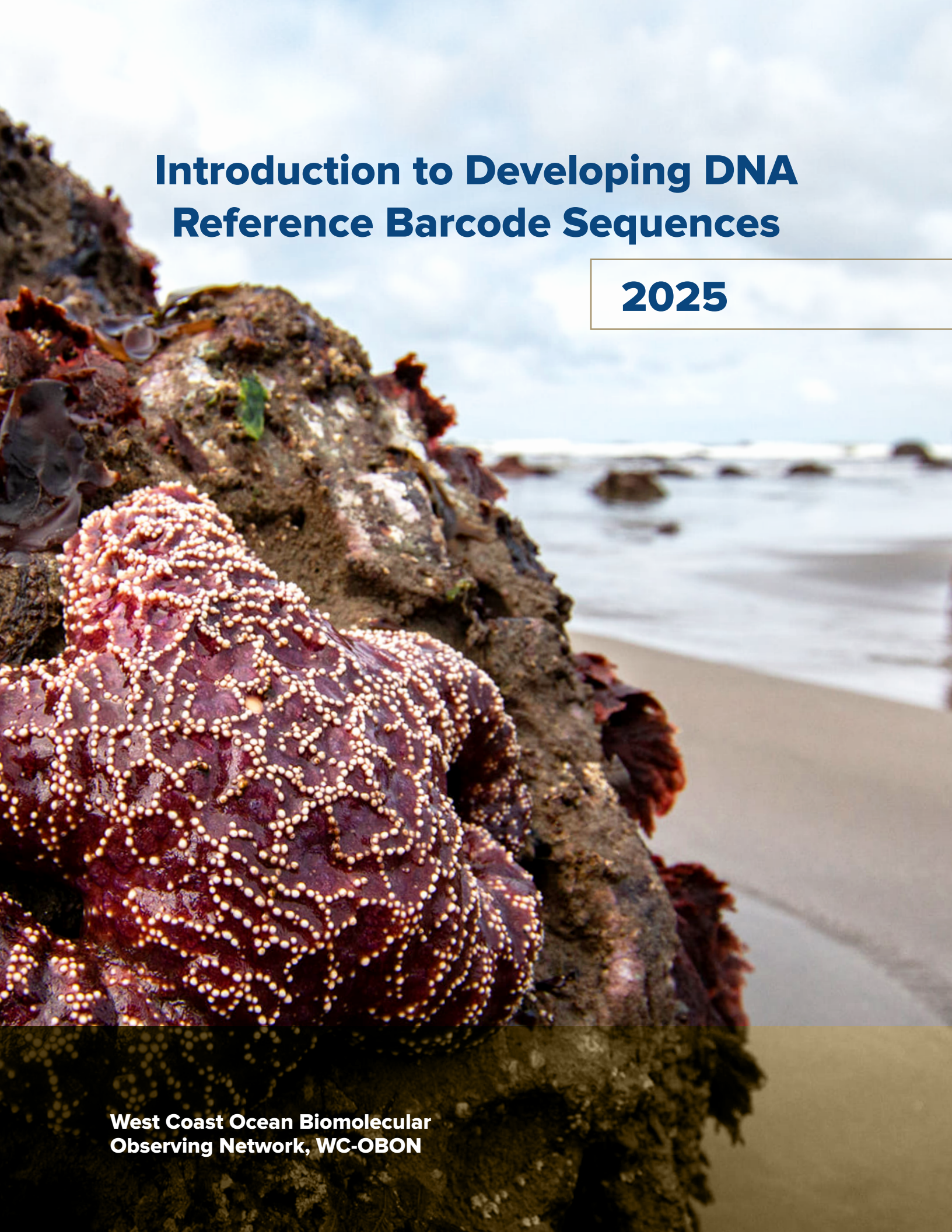
Introduction to Developing DNA Reference Barcode Sequences

2025



West Coast Ocean Biomolecular Observing Network, WC-OBON

This document was authored by the West Coast Ocean Biomolecular Observing Network (WC-OBON), a project under the Ocean Biomolecular Observing Network Programme (OBON), sponsored by the United Nations Decade of Ocean Science for Sustainable Development 2021-2030 (U.N. Oceans Decade).

DOI # 10.5281/zenodo.14867763

Preferred citation format: The West Coast Ocean Biomolecular Observing Network (2025). Introduction to Developing DNA Reference Barcode Sequences. 10.5281/zenodo.14867763

West Coast OBON OCEAN BIOMOLECULAR OBSERVING NETWORK

https://evsatt.github.io/WC-OBON_Website/

2021 2030 United Nations Decade of Ocean Science for Sustainable Development

https://forum.oceandecade.org/

Cover: Ochre star (*Pisaster ochraceus*) in Olympic Coast National Marine Sanctuary, Washington. © Zack Gold

February 2025 | **NOAA PMEL #2025-029_v3.2** | Layout and design by Sarah Battle

Contributors *(in alphabetical order)*:

Shannon Brown
University of Washington Cooperative Institute for Climate, Ocean, and Ecosystem Studies (UW CICOES) and National Oceanic and Atmospheric Administration (NOAA) Pacific Marine Environmental Laboratory (PMEL)

Allen Collins
NOAA Fisheries Office of Science and Technology (OST) and Smithsonian National Museum of Natural History (NMNH)

Matthew Girard
Smithsonian National Museum of Natural History (NMNH)

Zachary Gold
National Oceanic and Atmospheric Administration (NOAA) Pacific Marine Environmental Laboratory (PMEL)

Kelly Goodwin
National Oceanic and Atmospheric Administration NOAA Ocean Exploration

Sean M McAllister
University of Washington Cooperative Institute for Climate, Ocean, and Ecosystem Studies (UW CICOES) and National Oceanic and Atmospheric Administration (NOAA) Pacific Marine Environmental Laboratory (PMEL)

Christopher Meyer
Smithsonian National Museum of Natural History (NMNH)

Kim Parsons
National Oceanic and Atmospheric Administration (NOAA) Northwest Fisheries Science Center (NWFSC)

Nastassia Patin
California Cooperative Oceanic Fisheries Investigations (CalCOFI) and Southern California Coastal Water Research Project (SCCWRP)

Gregory Rouse
UC San Diego's Scripps Institution of Oceanography (UCSD SIO) and Benthic Invertebrate Collection (BIC)

Linsey Sala
UC San Diego's Scripps Institution of Oceanography (UCSD SIO) and Benthic Invertebrate Collection (BIC) and Pelagic Invertebrate Collection (PIC)

Erin Satterthwaite
California Cooperative Oceanic Fisheries Investigations (CalCOFI), California Sea Grant, and UC San Diego's Scripps Institution of Oceanography (UCSD SIO)

Charlotte Seid
UC San Diego's Scripps Institution of Oceanography (UCSD SIO) and Benthic Invertebrate Collection (BIC)

Susanna Theroux
Southern California Coastal Water Research Project (SCCWRP)

Regina Wetzer
Natural History Museum of Los Angeles (NHMLA)

Bat Star and *Dictyota* brown algae off Anacapa, CA. © Zack Gold

Contents

Detailed Guide to DNA Reference Barcode Sequencing

Objectives and Motivation

This guide provides a framework for generating and disseminating voucher-based DNA reference barcode sequences. It provides a general step-by-step approach to collecting and processing specimens/vouchers, as well as generating and reporting the resulting nucleotide sequence data.

We describe protocols for molecular lab work for generating both reference gene sequences ("barcodes") as well as genome skimming to derive full mitogenomes and nuclear ribosomal repeat regions ("ultra-barcodes"). Our guiding principles include FAIR (Findable, Accessible, Interoperable, Reusable) and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics) data practices to ensure resulting specimen and sequence data will be publicly accessible.

We intend these guidelines to be useful for both novice and expert systematists and molecular biologists alike. Here we focus our examples on marine taxa and habitats, but this resource is widely applicable to most aquatic and terrestrial biodiversity and environments. We invite interested readers to study the Resources Section for a more in-depth treatment of the subject matter.

We intend these guidelines to be useful for both novice and expert systematists and molecular biologists alike.

Here we focus our examples on marine taxa and habitats, but this resource is widely applicable to most aquatic and terrestrial biodiversity and environments.

This work is led by the West Coast Ocean Biomolecular Observing Network (WC-OBON), a project under the Ocean Biomolecular Observing Network Programme (OBON), sponsored by the United Nations Decade of Ocean Science for Sustainable Development 2021-2030 (U.N. Oceans Decade).

Introduction

DNA reference libraries are an integral part of sequence-based biological monitoring. Reference sequences, such as those in GenBank and Barcode of Life Data System (BOLD), are most useful for identification of species from environmental samples when they are based on publicly accessible, taxonomically identifiable morphological specimens.

To generate DNA reference sequences, voucher specimens are collected from the environment, taxonomically identified by an expert, accessioned, and cataloged into a long-term, publicly-available, curated museum collection (Figure 1). This practice ensures current and future researchers may obtain well-preserved specimens for examination and further study, including comparative and new species descriptions and interrogation with emerging technologies.

Field expedition efforts should coordinate with intended holding institutions (e.g. natural history museums) prior to the initiation of any collecting activities in order to generate a mutually agreed upon sample management plan. These plans help museums and collectors coordinate expectations and project required costs, space, and personnel. Tissue samples from vouchers can be sub-sampled in either the field or the museum and extracted in the lab for DNA sequencing.

DNA reference barcodes are short DNA sequences from standard genetic loci that are used to identify species. As sequencing technology has advanced, so have DNA reference barcodes. To date, the vast majority of reference DNA barcodes have been generated through capillary or "Sanger" sequencing, targeting short mitochondrial or nuclear ribosomal DNA sequences. However, as sequencing costs decline, the field is advancing toward genome skimming, which allows for the assembly of the entire mitochondrial genome, nuclear ribosomal repeat regions, and other nuclear gene regions, enabling the generation of reference barcodes for multiple genes ("ultra-barcodes") simultaneously (Figure 2).

Using vouchers for reference barcoding efforts is critical as species identifications are hypotheses and may change over time as we learn more about the taxonomic relationships among species. See Resources Section 6.

Scientists curate, organize, and update important taxonomic and biodiversity information progressively. Only if data connections are maintained between voucher specimens and their derived data in public repositories, can these taxonomic updates cascade to inform prior sequence identification annotations.

Recording accurate data and metadata and maintaining data linkages will increase the accuracy and reliability of DNA reference databases and molecular approaches like environmental DNA that rely on the most up-to-date reference materials.

Juvenile Puget Sound King Crab (*Echidnocerus cibarius*) off of San Juan Island, Washington. © Zack Gold

Summary of Recommendations

Task	Minimum Standard	Recommended Standard
Planning and Preparation	• Define objectives • Secure necessary permits • Generate collection data sheets	• Work with a permanent collection to delineate workflow and obtain established protocols • Adhere to Darwin Core standards
Field Collection	• Collect specimens in a way that ensures high quality DNA • Record accurate metadata • Label appropriately • Ensure safety of team	• Follow institutional guidance on specimen collection • Follow DwC and MiXS metadata standards
Live-Sorting and Pre-Processing	• Identify individuals suitable for DNA sequencing	• Identify specimens to morphospecies • Minimize harm • Minimize cross-contamination • Follow taxon-specific best practices for preservation
Photographing	• Take high quality images of specimens • Link photographs to metadata	• Include scale references • Include specimen codes in photographs
Tissue Sampling	• Collect suitable tissue and preserve for sequencing	• Preserve voucher specimens following taxon-specific best practices
DNA Extraction	• Use appropriate protocol in a dedicated molecular biology space • Quantify the DNA yield	• Archive DNA for long-term storage • Follow institutional best practices
DNA Barcode Sequencing	• Sanger sequence relevant marker or conduct genome skimming	• Shotgun sequence the entire mitochondrial genome and rRNA repeat regions
Bioinformatics	• Assemble sequences • Conduct QA/QC of resulting sequence data • Align sequences with existing references to confirm accuracy	• Generate alignment(s) and phylogeny(ies) of marker gene(s) to verify taxonomy • Follow FAIR and CARE principles for data and code accessibility and reproducibility
Archiving and Data Integration	• Submit raw reads to NCBI SRA • Submit sequence to NCBI GenBank • Archive sample in long term repository • Verify data submission	• Follow DwC and MiXS standards • Generate a BioProject • Archive biomaterial and associated data at partnering institutions

Table 1. Summary of minimum requirements and recommended best practices for Developing DNA Reference Barcode Sequences.

Importance of Developing DNA Reference Barcode Sequences Aligned with Darwin Core Standards

An essential overarching theme of developing DNA reference libraries is aligning efforts with the widely used biodiversity informatics standards to ensure that the data generated can be more easily integrated into broader marine biodiversity datasets. This is critical for contributing to larger scale, global ecological, conservation, and research efforts that require mobilization and sharing of biodiversity data for characterizing and understanding marine life.

Successful sharing and integration require robust biodiversity data standards to ensure interoperability and traceability of harmonized data. Accurate reporting of data and metadata at every step of the process, from field sampling, site and environmental information, methods and protocols, identifications of species, and associated measurements, are critical for linking the data together in ways that maximize data utility and usability.

Simply put, the Darwin Core Standard provides the data management structure to ensure harmonization of data practices across the biodiversity reporting landscape.

Darwin Core (DwC) is a standard maintained by TDWG (originally the Taxonomic Database Working Group, now the Biodiversity Information Standards) to facilitate the sharing of information about biological diversity. It includes a series of terms that define various attributes in order to allow interoperability and aggregation.

Importantly DwC is the standard used by the majority of biological collections and aggregators worldwide to ensure robust and FAIR data. DwC is based on a broad backbone of terms associated with events, occurrences, identifications, taxa, locations, record-level identifications, and measurements or facts (Figure 1).

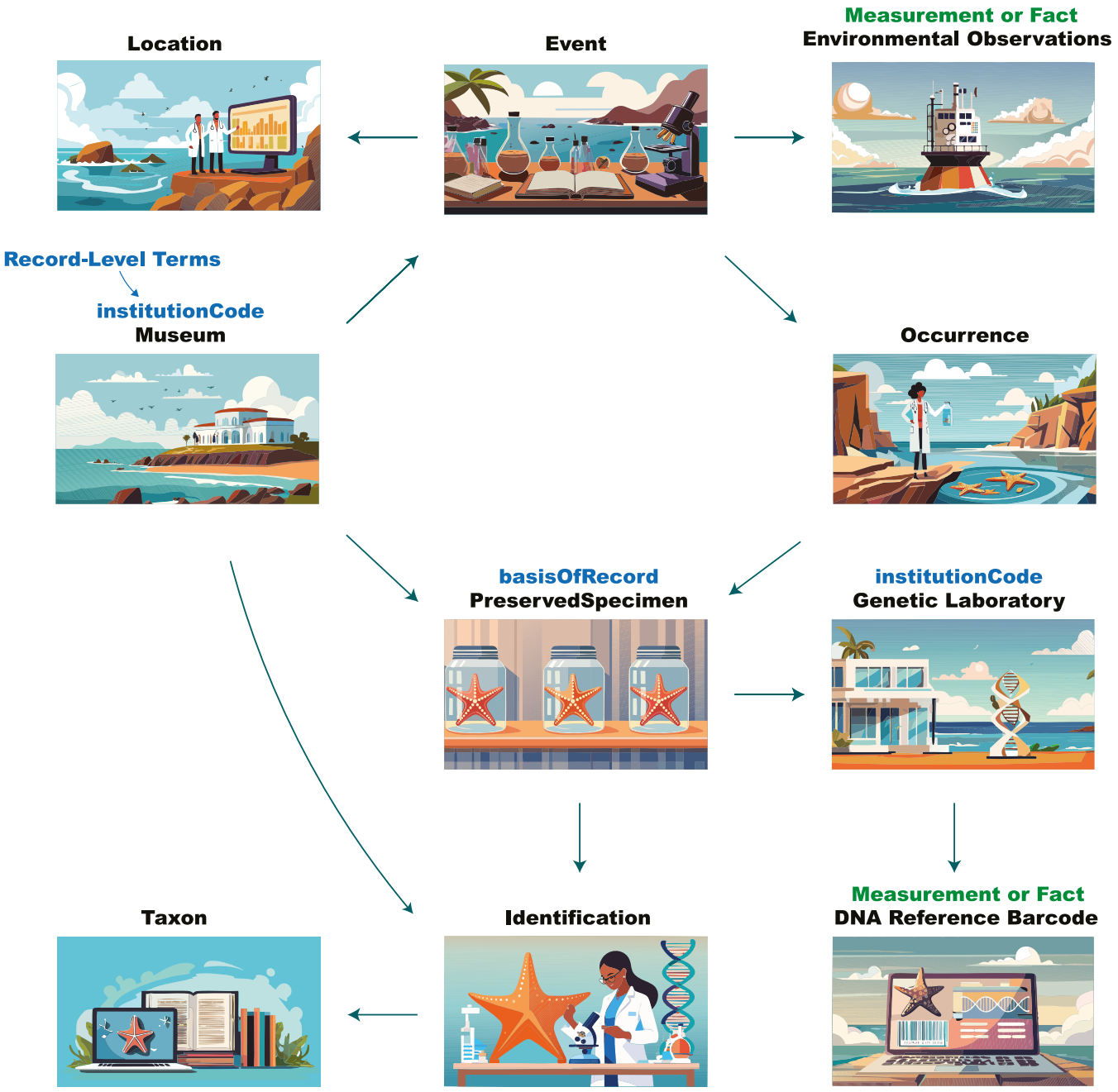


Figure 1. Key Darwin Core Standard Terms.

Here we illustrate how Darwin Core terms are used to link information about a specific observation effort (Event) to both the place (Location) and relevant geographic and biophysical data (Measurement or Fact). An event is conducted by an agent (e.g. individual, collector), resulting in an observation of an organism at a particular time and place (Occurrence). The organism can be collected and vouchered in a holding institution or permanent collection repository (e.g. museum, biorepository), the specific nature of this data record (basisofRecord) being a preserved specimen. This voucher is then assigned (Identification) to a lowest Taxon (e.g. specific species, genus, family). For DNA barcoding, the laboratory procedures (e.g., DNA extraction and sequencing) are linked to the institutions and facilities (Record-Level Terms) conducting the work. A researcher then processes a final reference sequence (Measurement or Fact) and uploads it to publicly available sequence and biodiversity data portals.

By adopting and linking this information, the Darwin Core architecture meets sample and data management needs, including ensuring both the chain of custody and metadata information is preserved and available to the user community. Simply put, the Darwin Core Standard provides the data management structure to ensure harmonization of data practices across the biodiversity reporting landscape.

The Darwin Core effort encompasses classical biodiversity observations with an expanding focus on digital and molecular observations. For reference barcoding, incorporating Darwin Core standards throughout the process ensures high-quality data associated with voucher specimens and reference sequences, increasing the value and reliability of data sources. Adoption of these data standards and best practices is required to ensure trusted and reliable genetic sequence data. Implementing this robust documentation and organization scheme is especially important for conservation and management decisions that rely on the accuracy of such DNA and biodiversity resources as highlighted in the **U.S. National Aquatic eDNA Strategy**.

Key Darwin Core Standard terms include the following:

- An **Event** is when and where observations of biodiversity took place and includes information on the sampling protocols and methods, date, time, and field notes. For example, collecting a sea star from a tide pool or a marine trawl at a specific time and place.

- **Occurrences** refer to an observation of an organism at a particular place and time. Occurrences are inherently associated with an event as well as with the identification and the described taxonomy of the organism. For example, an ochre sea star (*Pisaster ochraceus*) observed from the tide pool above (Figure 1) or a *Limacina helicina* pteropod collected in a bongo net tow.

- **basisOfRecord** refers to the specific nature of the data record whether it is an observation of a species made in the field or a preserved specimen vouchered in a museum.

- **Identification** is the taxonomic determination of an organism and serves as the linkage between occurrences and taxon categories. Terms associated with the identification can include the nature of the evidence (e.g., expert identification under a microscope or a cytochrome B sequence) and any associated identification qualifiers (e.g., *Hemisquilla* sp.).

- A **Taxon** is a group of organisms considered by taxonomists to form a homogeneous unit. This includes information on scientific names, vernacular (common) names, taxon concepts, and relationships between them. For example, *Phalacrocorax pelagicus* is the scientific name of the commonly known Pelagic Cormorant with Taxonomic Serial Number (TSN) 174725 (TSN is the unique identifier for each taxon in the Integrated Taxonomic Information System (ITIS)) and *Chalinidae* is the scientific name for a family sponges with AphiaID 131636 (AphiaID is the unique identifier for each taxon in the World Register of Marine Species (WoRMS) database).

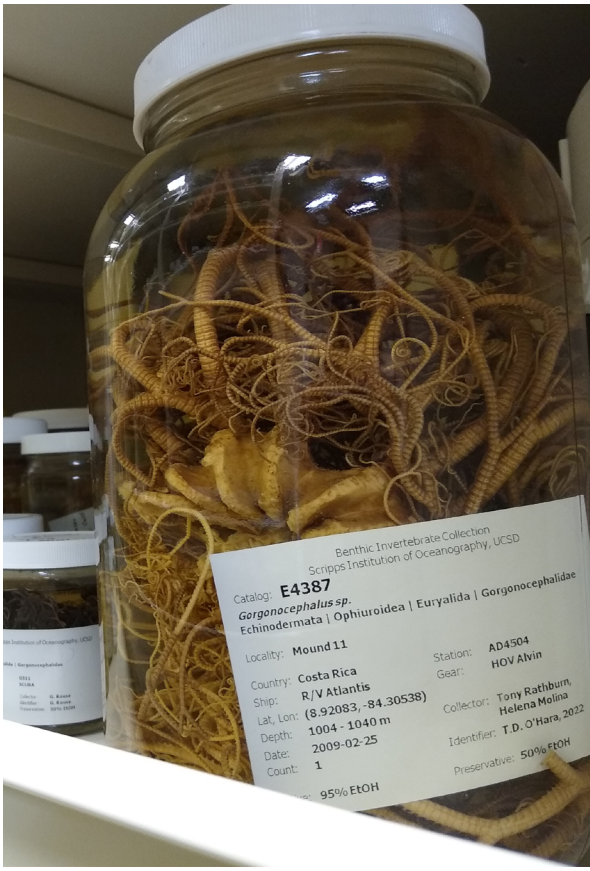
- The **Location** of an event describes the geography, locality descriptions, and spatial data. For example, the GPS coordinates (34.001348, -118.804790) of an intertidal quadrat at Point Dume State Marine Reserve.

- **Record-level terms** describe the who of an event or occurrence including detailed information on the associated catalog, collection, institution, scientist, and dataset identifications.

- **Institution Code** is a unique identifier for the permanent collection or facility having custody of the vouchers, tissues, sample materials, DNA extracts, sequence information, data, or metadata referred to in the record.

- **Measurement or Fact terms** are characteristics, assertions, or references associated with organisms, occurrences, events, identifications, or taxon. This is a broad-ranging category and includes the weight of organisms in grams, the number of arms of an organism, **Environmental Observations** (e.g. the surface water temperature where the organism was collected), and/or DNA reference barcode sequence.

The Global Biodiversity Information Facility (GBIF) and Ocean Biodiversity Information System (OBIS) are repositories for biological observations from both non-specimen and specimen based occurrence records, including those specimens used for barcoding. They have helpful resources for the use of Darwin Core terms and information when collecting specimens and submitting data. See Resources Section 3 for more information on Darwin Core.



Detailed record label of *Gorgonocephalus sp.* that includes key relevant Darwin Core Standard information. © Charlotte Seid

Detailed Guide to DNA Reference Barcode Sequencing

1. Planning and Preparation

Define Objectives

Determine the goals of the barcode sequencing project, including the geography, realm, target taxa, target gene(s), and specific research questions.

Establish where vouchers will be deposited (specific permanent, digitally discoverable collection) and coordinate institution-specific best practices. Here, we strongly encourage establishing a relationship with an existing permanent repository to accept voucher specimens and data prior to initiating the project.

Permanent collections bring decades to centuries of institutional knowledge on robust sample and specimen deposition protocols with collection-specific approved containers, labels, protocols, etc. and often have detailed resources available online or on request.

Permanent collections are museums or similar entities that have dedicated collections management staff to care for the material in the long term and are able to loan material to experts for further study. These repositories can help advise you on the multifaceted process of sample collection, packing and shipping of specimens, and vouchering, including navigating streamlining data and sample collection efforts. Importantly permanent collections bring decades to centuries of institutional knowledge on robust sample and specimen deposition protocols with collection-specific approved containers, labels, protocols, etc. and often have detailed resources available online or on request. See Resources Section for more information.

The majority of costs associated with a reference barcoding project occurs in the field collection, identification, and vouchering phases. Therefore it is critical to discuss funding for these efforts in the scoping and planning stages of the project. This is especially true for accessioning voucher specimens into a specific permanent collection as sample curation is labor intensive and requires dedicated supplies and infrastructure to sustainably manage deposited samples for decades to centuries.

Robust identification of specimens is also time consuming, requiring both morphological observation, sometimes through specialized techniques (e.g., scanning electron microscopy) and scholarship through reference to literature and other specimens. Funding discussions to ensure identification and vouchering should take place well in advance of project implementation.

Secure Permits and Permissions

Obtain necessary permits for collecting specimens, and ensure compliance with local, national, and international regulations. Most states and federal agencies as well as tribes, universities, and private property holders require permits for specimen collection. Protected species, sensitive habitats (e.g., California Marine Protected Areas (MPAs) or U.S. National Parks), and species covered by Institutional Animal Care and Use Committees (IACUC) have their own specific additional permitting requirements. The onus of complying with such requirements is on the individuals conducting the collection. The permitting process may take months, so it is important to plan accordingly and start the process early.

Again, we strongly recommend coordinating with staff at existing permanent collections where samples will be archived. Permanent collections may have specific requirements with regards to accepting samples and ensuring chain of custody under a specific permit and capacity for reuse. Furthermore, museum staff are often knowledgeable on relevant permits for sample collection and can advise on the permitting process to ensure compliance with relevant laws and jurisdictions.

Organize Equipment and Supplies

Gather all required equipment, such as field collection tools, preservation chemicals, and laboratory supplies. Follow guidelines for any shipping restrictions on hazardous chemicals such as ethanol. For example, there are specific trainings, rules, and regulations for shipment of specimens/hazardous goods (See Resources Section 6).

Establish Protocols

Develop or obtain standardized protocols for field collection and documentation, specimen handling, preservation, DNA extraction, sequencing, and data analysis and management. Ensure these protocols are shared with and agreed upon with staff at permanent collection prior to an expedition. This guide and the Resources Section can help direct users in the appropriate direction.

Follow Darwin Core Standards

Follow Darwin Core (DwC) standards for recording metadata from biodiversity sample collections. Here we emphasize the importance of ensuring high quality record keeping for priority metadata: eventDate (e.g., the day, month, year, and time of the sampling event in Coordinated Universal Time - UTC), decimalLongitude and decimalLatitude (e.g., the GPS coordinates of sampling event), depth (e.g., vertical location in the water column).

We highlight that OBIS currently requires seven and strongly recommends one DwC terms: occurrenceID, eventDate, decimalLongitude, decimalLatitude, scientificName, occurrenceStatus, and basisOfRecord, with scientificNameID strongly recommended.

Specific metadata and environmental data may be a priority for your field or required by your funder. We strongly encourage complying with ongoing relevant national and international biodiversity, molecular, and ocean observing standards consortia to ensure the adoption of best practices. See Resources Section 4 for a collated list of consortia and standards.

Timely and accurate record keeping is the priority. Specific metadata and data formatting can be adjusted later from high quality source data if needed. See Resources Section 5 for detailed guidelines and recommendations for record keeping.

Generate a Workflow for Sequencing Based on the Goals of the Study

Plan the technical workflow based on study goals. For example: Can the specimen be identified at the species level based on morphology? Is/are the target organism(s) well studied with one or more existing marker genes in public databases? Is the goal to design new eDNA assays for target species? Is the goal to fill out existing reference databases for an existing eDNA assay? Are you planning on generating a robust phylogeny with multiple sister species and/or clades? The answers to such questions will determine the workflow and ensure that work proceeds efficiently and effectively.

The folowing terms are strongly recommended:

occurrenceID
eventDate
decimalLongitude
decimalLatitude
scientificName
occurrenceStatus
basisOfRecord with scientificNameID

See the Additional Considerations (DNA Sequencing) Section to guide your workflow selection between Sanger sequencing and genome skimming.

2. Field Collection

Collect Specimens

Collect specimens as appropriate to keep them alive and intact for further sorting, processing, and photography. For example, maintain temperature, oxygen, and water levels. DNA degrades after cell death, so live organisms with fresh tissue provide the best starting material for reference barcoding and species identification efforts.

Some species can only be identified morphologically when fresh (e.g., gelatinous invertebrates, shrimp) whereas others often can be identified after preservation (e.g., fishes). Life color and delicate features often degrade in stressed organisms and these features may be diagnostic characters necessary for distinguishing species, requiring photography of live specimens and care to preserve high quality specimens. This will affect where and when taxonomic expertise is required (e.g., taxonomists leading sample collection efforts or receiving processed samples back in the laboratory).

Many vouchers should be preserved in formalin for further traditional taxonomic work (e.g. annelids, other soft bodied phyla). A common workflow for these species is to subsample tissues in the field from live animals into DNA-safe fixation (e.g. ethanol, flash frozen), relax the animal (for example using MgCl2, menthol, etc.), and then preserve the organism in formalin to allow for the phenotype to be appropriately preserved for future morphological investigation and refinement of identification.

Top: Intertidal field collections on Catalina Island led by Natural History Museum of Los Angeles County scientists. © NHMLAC

Bottom: Dr. Regina Wetzer identifying a specimen collected from the field. © NHMLAC

Record Metadata

Document comprehensive metadata for each specimen, including (but not limited to) date, time, location (latitude and longitude), collectors, collection method, depth, locality name, habitat type, and environmental conditions. Follow DwC and MIxS (Minimum Information about any (X) Sequence) standards for determining required information. See Resources Sections 3 and 4 for more information. When working with a repository, it is important that you are aware of institution-specific DwC data requirements.

Label Specimens

Use legible, durable, and ethanol-resistant labels with unique alphanumeric identifiers (that is, unique identification codes that unambiguously connect a physical specimen to its associated data; consult partner collections for institution-specific guidance). Ensure that all collected specimens are labeled accurately in such a way that could be interpreted by anyone at any point in time without specific knowledge about that collection event, specimen, location, etc. We advise developing best labelling practices in coordination with the museum that will house voucher specimens.

Well-labeled marine invertebrate sample vouchers in the UCSD Scripps Institution of Oceanography Benthic Invertebrate Collections. © Charlotte Seid

3. Live-Sorting and Pre-Processing

Sort Specimens

Sort live specimens by morphospecies (i.e. physical characteristics, or morphology). Use relaxing agents (e.g., magnesium chloride, clove oil, or menthol) if necessary to minimize damage to the animals, treat them humanely in accordance with approved IACUC protocols, and minimize handling for photography. Some organisms require relaxing for photography, but also to minimize DNA degradation. It is also important to ensure any relaxing agent does not impair DNA recovery (e.g. clove oil can degrade DNA at sufficiently high concentrations).

Minimizing human or cross-species contamination is ideal (contamination here is defined as a non-target species DNA unintentionally found in results), particularly when the organism of interest is small. Best practices include use of gloves, careful removal of epibionts, maintaining water levels, oxygenation, and temperature, avoiding excessive cross-contamination of water and chemicals, and separating specimens early while they are alive.

See Resources Sections 2.2 and 7 for more detailed information.

Assess Size, Variation, and Condition

Evaluate the size, variation, and condition of specimens to decide which samples are suitable for DNA extraction and barcode sequencing. Larger specimens are more amenable to DNA extraction from a subsample while preserving other parts of the specimen. Smaller specimens will require different care, handling, preservation, and potential dissection preparations, and may require microscope slides. Regardless of size, an intact organism in good condition will increase the success of accurate identification and sample preservation.

4. Photographing

Above: Santiago Herrera captures high resolution photographs of a cold-water coral specimen from the Gulf of Mexico. © *Lophelia II 2009: Deepwater Coral Expedition: Reefs, Rigs and Wrecks* sponsored by the NOAA Office of Ocean Exploration and Research (OER)

Next page: Collage of high resolution invertebrates from the California Current: 1) *Hermissenda opalescens nudibranch*, 2) *Limaria hemphili* file shell, 3) *Neanthes sp.* polychaete worm, 4) *Ophiuroidea* brittle star, 5) *Leptopecten latiauratus* kelp scallop, 6) *Munidopsis girguisi* squat lobster, 7) *Poraniopsis inflata* sea star, 8) Fionoidea nudibranch, 9) *Astrodia sp.* brittle star on *Swiftia sp.* Gorgonian. © Charlotte Seid

Capture Images

Take high-quality photographs of specimens, using a microscope and camera adapter where appropriate, to document physical characteristics including aboral/oral surfaces, anterior/posterior ends, and any unique features, especially those relevant for identification purposes. For gelatinous organisms, internal canals are often important for identification and can be visualized and photographed under strong side lighting. Color pattern is also important to capture as preservation often removes coloration. Ensure images are clear and include scale references (e.g., a ruler or the magnification used with a photomicroscope, or a standardized font specimen label with an underscore can be used for scale).

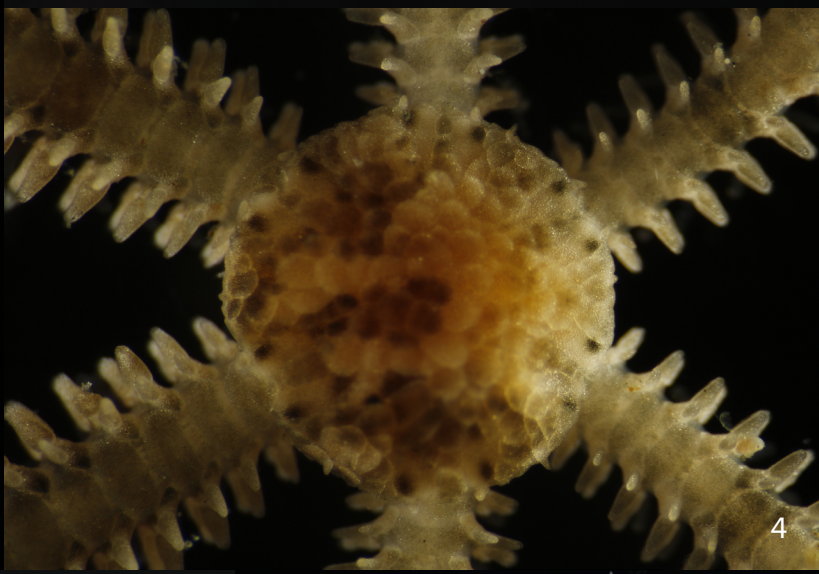
Record Unique Codes

Include alphanumeric identifiers that are unique for each specimen in photos where practical to ensure that these identifiers appropriately link data streams.


1


2


3


4


5


6


7


8


9

5. Tissue Sampling and Specimen Preservation

Sampling Small Organisms

When an organism is very small (< ~1 mm in diameter), most likely it will be necessary to sample the entire individual for DNA extraction. In these cases, double check that appropriate photographs have been taken and/or preserve additional specimens of the same morphotype (sometimes referred to as paravouchers) if possible that are clearly linked informatically to the consumed individual.

Subsample Tissue of Moderate to Large Organisms for DNA Reference Barcoding

Carefully subsample tissue for DNA extraction. Avoid the digestive tract and select DNA-rich tissues. See Resources Sections 2,6, and 7 for taxon-specific guidance of an appropriate size ("if you can see it, you can sequence it"). Be careful not to cross-contaminate with other specimens (e.g., separate specimens and sterilize sampling tools when possible).

Preserve the Tissue Subsample for DNA Reference Barcoding

Preserve the sample as appropriate for subsequent DNA extraction, i.e., in 95% ethanol, cold or frozen (flash frozen in liquid nitrogen or in -80˚C preferred), and keep the samples in the dark, if possible. We recommend a minimum of 1:10 sample:ethanol ratio. The original ethanol is diluted by water from the sample's cells and the sample risks degradation if the ethanol concentration is not maintained near 95%. Therefore, it is often necessary to exchange the ethanol preservative approximately 24 hours after initial fixation and again as needed. Keep the samples cold and in the dark if possible for long-term storage.

Relax, Fix and Preserve Voucher Specimen

Relax, fix and preserve remaining specimens as appropriate for future morphological study and/or DNA sequencing (if preserved in ethanol). Relaxation techniques are taxon specific and can be found in the various references. Fixation for morphological work for many marine invertebrates can require fixation in 10% formalin (=3.7% formaldehyde in seawater - buffered if preserving in formalin) or in 95% ethanol, depending on the taxon.

Preservation for Reference DNA Barcoding	Relaxation	Fix	Long-term Preservation
95% ethanol, cold or frozen (flash frozen in liquid nitrogen or in -80˚C preferred), and keep the samples in the dark, if possible There should be a minimum of 1:10 sample:ethanol ratio	Techniques are taxon specific and can be found in various references	10% formalin (=3.7% formaldehyde in seawater - buffered if preserving in formalin) or in 95% ethanol, depending on the taxon	Some should be stored in formalin (many cnidarians, tunicates), others (most) should be rinsed in **fresh** water and then gradually transitioned into ethanol after appropriate fixation in formalin (days to months, depending on size) There should be a minimum 5:1 fluid:specimen ratio

Table 2. Specimen preservation recommendations (combine with taxon reference).

Long-term preservation fluid varies by taxon. Some should be stored in formalin (many cnidarians, tunicates), others (most) should be rinsed in **fresh** water and then gradually transitioned into ethanol after appropriate fixation in formalin (days to months, depending on size). Formalin forms cross-links between DNA and proteins, causing single-strand breaks and conformation changes in DNA, and thus is challenging for downstream DNA applications. See Resources for more information.

There should be a minimum 5:1 fluid:specimen ratio. When using 95% ethanol to preserve specimens, replace the preservative with fresh ethanol 24 hours after initial treatment. Repeat ethanol replacement as needed (i.e., replace ethanol if preservative turns amber because of leached pigmentation) to account for dilution of preservative fluid. Consult the recipient collection for appropriate shipping instructions and archival containers and the desired final concentration of preservation fluids.

Clearly label what chemical preservative is used in each container and ensure hazardous materials are appropriately labeled.

See Resources Sections 2 and 6 for more detailed information.

6. DNA Extraction

Extraction

We provide a brief overview of DNA extraction methodologies as they have been well documented elsewhere. See Resources Section 7 for more detailed information.

Most tissues can undergo a standard DNA tissue extraction method (e.g., Phenol:Chloroform, Qiagen DNEasy Blood and Tissue Kit, Zymo or similar) to purify DNA. However, some taxonomic groups including Nemertea, gastropod molluscs, ophiuroid and crinoid echinoderms, and some crustaceans, particularly individuals with bright color pigments and/or those that produce a lot of mucus may have improved DNA extractions with more complex protocols that remove inhibitors and other biological compounds are necessary (e.g., Qiagen DNeasy Plant or E.Z.N.A. Mollusc and Insect DNA Kit, etc.).

Use sterile technique with molecular grade reagents in all cases to avoid contamination of DNA from sources outside of your sample. Quantify the resulting DNA yield using a fluorometry based approach (e.g., Qubit).

DNA Sequencing Considerations

Here we discuss Sanger sequencing and genome skimming workflows and which approach is best suited for given research questions and project objectives.

Sanger sequencing is a widely used, comparatively simple sequencing approach targeting single marker gene sequences (Figure 2, Table 3).

Genome skimming is a more advanced sequencing approach targeting the entire mitochondrial genome and can recover nuclear ribosomal repeat regions and other genes as well. In some cases, it may be advantageous to screen a set of specimens using Sanger sequencing of marker genes before selecting the most appropriate samples for genome skimming.

We provide a series of situations and recommendations for deciding between Sanger sequencing and genome skimming work flows (Table 4), as well as more details on both methods in the subsequent sections.

Sanger Sequencing

Sanger sequencing of a single reference barcode (targeting 200–1000 base pairs in length, typically sequenced in both directions) is generally appropriate for assigning **individual specimens** to distinct "species" categories (species delimitation). However the efficacy of species delimitation for a single reference barcode depends on the rates of evolution in a lineage and genetic marker. Thus it may not be sufficient for full taxonomic identification, requiring further analysis of more DNA sequences (e.g. additional Sanger-derived sequences or genome skimming) (Figure 2).

There are several advantages of Sanger sequencing for reference barcode generation: 1) lower cost, 2) lower barrier to entry of bioinformatics skills needed, and 3) decades of previous workflows and protocols upon which to build. Sanger sequencing is a tried and true approach to generating reference barcodes and has been the staple of DNA sequencing for over three decades, resulting in millions of deposited sequences in public sequence repositories.

Sanger sequencing does have some disadvantages and challenges including 1) the potential for poor alignment between the target sequence and commonly used primers making it difficult to obtain an amplicon, and 2) contamination of samples by bacteria or other organisms such as gut contents or endo/ecto-symbionts resulting in an unreadable/ambiguous result. Primer misalignment can be mitigated by redesigning primers (time consuming) or choosing alternative reference barcode regions. Contamination of samples can be mitigated by proper sterile technique and targeting appropriate tissues with fewer symbionts/potential for contamination. In contrast, genome skimming does not rely on primers so avoids misalignment issues, and the high sensitivity and sequencing coverage (e.g. the number of unique sequencing reads that align to a specific region) increases the viability of generating mitogenomes from contaminated samples.

Sanger sequencing advantages for reference barcode generation:

1) lower cost

2) lower barrier to entry of bioinformatics skills needed

3) decades of previous workflows and protocols upon which to build

Genome Skimming

Genome skimming is generally appropriate for one to a few representatives per species to provide both a complete mitogenome reference and nuclear ribosomal repeat regions (Figure 2).

Genome skimming targets ~20-40X coverage of the genetic data for a single marker gene. Its scientific benefit is derived from the fact that it generates reference data for all high-copy genetic regions that include likely candidate marker genes from both organellar (mitochondial DNA (mtDNA) and chloroplast DNA (cpDNA)) genomes and common nuclear repeat elements. The ability to compare different marker genes and not be tied to existing household regions opens the door to designing species- or clade-specific assays, generating robust phylogenies, and assessing rates of evolution in ways that are not possible with single marker genes. Genome skimming can also reveal the identity of symbiont taxa, for instance the rich microbial communities living in association with sponges. The additional cost and labor can be a worthwhile tradeoff to provide a more comprehensive genetic resource that enables a broader range of science than Sanger sequencing and is typically more informative for eDNA assay development and population genetics.

Genome skimming advantages:

1) generates more comprehensive genetic resource

2) applicable to broader array of scientific research questions including: eDNA, robust phylogenies, and symbiont taxa

Genome skimming is a more complex process compared to single gene Sanger sequencing, particularly regarding the need for bioinformatic expertise to assemble high throughput DNA sequences. We note that the increased bioinformatics processing increases the overall time and cost of genome skimming, but is highly dependent on the expertise of the researcher(s) and availability of existing turnkey pipelines

Increasingly, genome skimming of voucher specimens is being used to increase the value of such data for a broad range of research, monitoring, management, and decision-making applications.

We also highlight that there are a myriad of applications of whole genome sequencing (Table 5). Whole genome sequencing is similar to genome skimming, but the aim is to assemble a more complete nuclear genome, requiring orders of magnitude more sequencing depth and longer reads, therefore significantly increasing the cost to ensure similar sequencing coverage for low copy gene regions (Tables 6 and 7).

Whole genome sequencing provides robust data on population connectivity, stock assignment, individual identification, evolutionary histories, and demographic histories. This work is outside the scope of this guide, but (See Resources Section 7) for more information.

Ultimately, it is up to the researcher and/or funding agency to determine which workflow serves the goals and objectives of the project best. Genome skimming to derive mitogenomes and ribosomal repeat regions provides more robust genetic resources at higher cost while Sanger Sequencing is simpler, faster, and more cost effective on a per sample basis.

Need for High Quality DNA

Regardless of approach, high quality DNA increases the success for DNA reference barcoding approaches. Degraded DNA, particularly from older and/or poorly preserved samples, needs more care and alternative preparation methods (e.g., formalin preserved samples).

Next generation sequencing approaches like genome skimming are typically more powerful than Sanger sequencing, given the high-throughput capacity of sequencing approaches, generating orders of magnitude more sequences (Figure 2). Genome skimming efforts can be advantageous for obtaining usable data from archived samples because they can recover smaller fragments and stitch them together informatically.

In other instances some samples may have higher success with Sanger sequencing given the nature of PCR amplification generating billions of copies of DNA from a few starting target sequences, enabling recovery of viable DNA sequences from otherwise low quality samples. In these cases, the researcher has to know what they are looking for and have specific targeted primers.

Genome skimming to derive mitogenomes and ribosomal repeat regions provides more robust genetic resources at higher cost while Sanger Sequencing is simpler, faster, and more cost effective on a per sample basis.

7. Sanger Sequencing

A critical consideration of Sanger sequencing as opposed to mitogenome skimming is the selection of the target marker to sequence. The chosen marker should match the intended application and care should be taken to identify marker gene targets and protocols that align for reference DNA barcode application.

A region of CO1 (mitochondrial cytochrome oxidase subunit 1, also known as COX1 or COI) is the most commonly sequenced reference barcode for marine metazoans given Barcode of Life efforts (Table 3; See Resources Section 7). However, the CO1 gene is not informative for resolving all marine species (e.g., corals, sponges, some abalone in the genus Haliotis, and any recent adaptive radiation) and thus other markers are needed.

Example Target Taxonomic Group	Commonly Used Marker Gene Set
Porifera	CO1, 28S rRNA, ITS rRNA
Cnidaria	CO1, mUTS, ITS rRNA, 16S rRNA, 28S rRNA
Arthropoda	CO1, 16S rRNA, 28S rRNA, 12S rRNA
Echinodermata	CO1, 12S rRNA, 16S rRNA
Vertebrates	CO1, 16S rRNA, control region (dLoop), cytB, 12S rRNA
Mollusks	CO1, 16S rRNA, cytB

Table 3. Examples of common marker genes for different taxonomic groups.

Sequencing Preparation

Using validated primers for commonly used marker genes (e.g., CO1, 16S, 12S, etc.), PCR amplify the gene and verify the amplification results using gel electrophoresis. Purify the PCR product. See Resources Section 7 for more detailed information on markers and methods.

After initial amplification and purification, conduct Sanger sequencing via chain-termination PCR followed by fragment visualization (Figure 2). If outsourcing to a sequencing facility, provide your generated amplicon and primer sequences for Sanger sequencing following their instructions. We strongly encourage that reference barcode sequences are generated from both forward and reverse primers to provide at least 2x coverage and correct for declining quality scores at the 3' end of each sequencing read.

Bioinformatics of Sanger Sequencing

Capture PCR primers and conditions in a standardized system of record, a Laboratory Information Management System, to ensure that downstream sequences are associated with the processes and individual that generated them. Verify the provided sequence data is in the correct file formats with associated sample metadata (e.g., fastq/fasta files or ab1 chromatograms). See Resources Section 5 for more information.

Sanger Sequencing Workflow

- Conduct quality control of the supplied chromatograms either by visually checking the sequence or by using the quality-informed output from the sequencing facility. It is critical to trim the low quality 3' end of each read before attempting overlap consensus (merged 2X sequence formation).

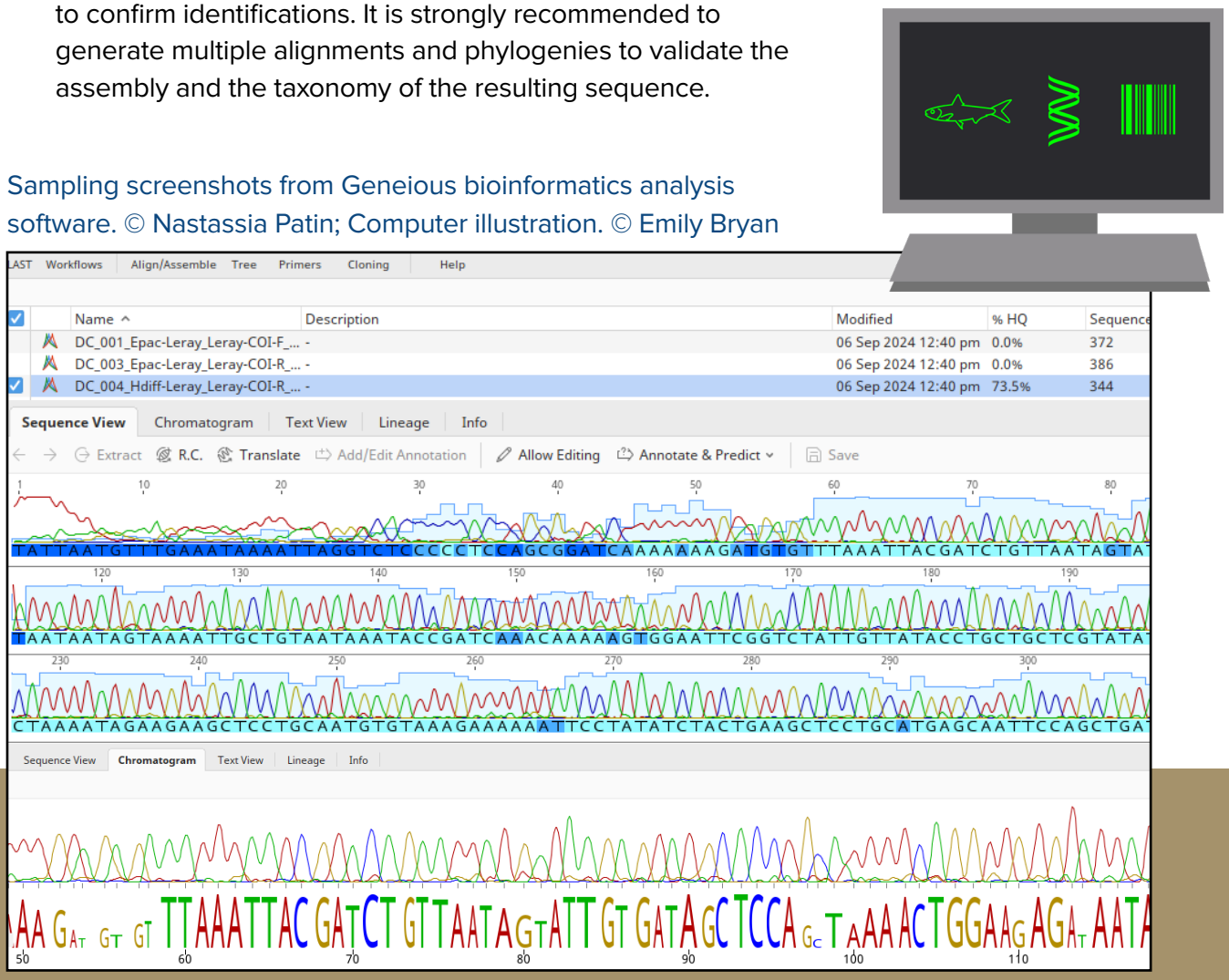
- Align the forward and reverse read to form the final 2x contig.

- Many researchers trim primers from the ends of the amplicon product. This is in part due to problems of quality at the 5' end of the sequencing read. In addition, this is in part due to some primers being intentionally degenerate (mixtures of similar primer sequences that incorporate variations at specific positions) and therefore the primer sequence may not be accurate to the organism sequenced.

- Both of these processing steps can be done using standardized and validated software (e.g., Geneious).

Notes on Quality Control/Quality Assurance

- Many ambiguous bases in the sequence or final contig are a sign of sample contamination, poor sequencing quality, or poor trimming. It may be necessary to trim additional bases to remove these degenerate bases.
- Poor sequence quality of the entire read cannot be bioinformatically resolved; either attempt additional sequencing or switch to genome skimming.
- Record if any chromatogram editing was performed and maintain a record of processes used to assemble DNA contigs.
- Convert the chromatogram to a fasta file and use the fasta file for subsequent data analysis and archival.
- Compare your sequences with existing reference libraries to confirm identifications. It is strongly recommended to generate multiple alignments and phylogenies to validate the assembly and the taxonomy of the resulting sequence.



Sampling screenshots from Geneious bioinformatics analysis software. © Nastassia Patin; Computer illustration. © Emily Bryan

8. Genome Skimming

Overview of Benefits/Costs

For genome skimming sequencing, the goal is to obtain at least 30-40x sequence depth coverage for the mitochondrial genome and ribosomal RNA repeat regions, typically aiming for **at least 3 gigabases (Gb)** of overall sequencing depth. The proportion of mitogenome and nuclear rRNA repeats to the rest of the nuclear genome varies with different taxonomic groups and tissue type, though it is generally 10s-1000s x more abundant.

The desired depth of sequencing also depends on the lengths of the target organism's nuclear genome and mitogenome which can also vary widely among taxa (expect > 10 kbp mitogenome and > 1 Gbp nuclear genome for most eukaryotes). If you can estimate the nuclear genome size of an organism, then a safe assumption is to plan on ~1% of sequence reads to be mitochondrial. However, determining the exact sequencing depth to ensure accurate and high-resolution mitogenomes is difficult and researchers may want to hedge their bets with more sequencing effort. For example, some researchers prefer to aim for 100x coverage and 15 Gb of overall sequencing depth to increase the likelihood of successfully reconstructing a high quality mitochondrial genome from skimming.

Sequencing Preparation

Identify a respected facility for shotgun sequencing of the organism's genome. Obtain and meet any quality control instructions or other guidance for sample preparation from the intended facility. Decide on the budget of the proposed project and whether the sequencing facility or researcher will be preparing the DNA for sequencing.

Some small organisms (< 1mm) may not yield enough DNA for genome skimming, and further representative whole-genome amplification may be required. For example, the GenomiPhi DNA Amplification Kit offers a highly efficient and representative whole-genome amplification.

If the researcher is preparing the genomic DNA extraction for sequencing, follow sequencing facility instructions including valid quality control/ quality assurance checks (e.g., quantification, size selection).

Bioinformatics of Genome Skims

We note that bioinformatics is an evolving field with new tools, approaches, and best practices continually emerging. We present the following information as an introduction to some of the bioinformatics tools commonly used as of February 2025, acknowledging that additional and newer approaches may be available.

Verify the provided sequence data are in the correct file formats (i.e. fastq files) with associated sample metadata.

Mitogenome Workflow

- Conduct quality control trimming of sequencing reads to remove adaptor and trim sequences or sequence ends with poor quality scores (i.e., <Q30). Failure to remove adaptors will lead to poor assembly results.

- Assemble reads into contigs using standard approaches (e.g., SPAdes, getOrganelle, mtGrasp, PMAT).

    - Often researchers first map reads (align) to a reference mitogenome to exclude nuclear genes and improve the accuracy and efficiency of assembly. Read mappers include minimap2 and Magic-BLAST, and mapping is the default approach in getOrganelle.

- Annotate the resulting mitogenome assembly with standard tools (i.e., mtGrasp, MITOS2, MitoZ).

- Pull out nuclear ribosomal repeat regions (e.g., SPAdes, getOrganelle, PhyloFlash)

Notes on Quality Control/Quality Assurance

- Efforts should be taken to circularize the mitogenome (if mitogenome is circular in species of interest) as well as confirm sufficient sequence coverage across all regions.

- After annotation is completed, manual curation will likely be necessary to confirm all expected genes have been annotated, that their start and stop codons are correct, and that genes are not split into multiple annotations.

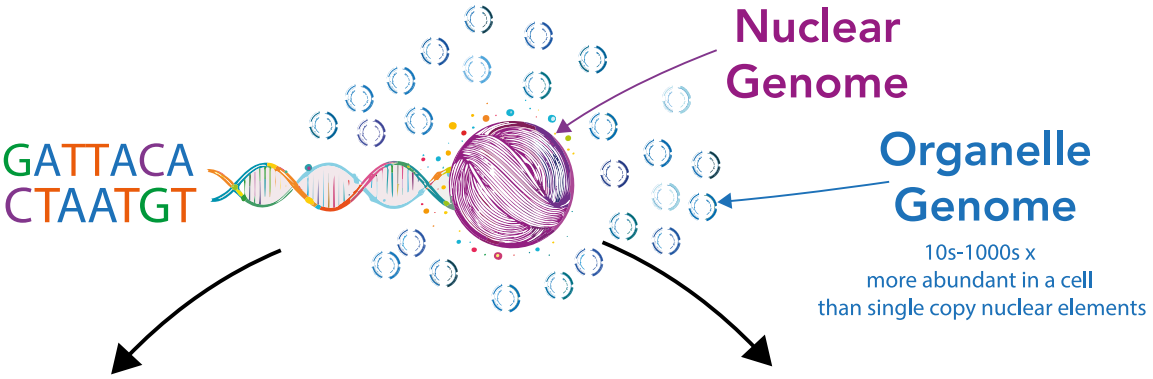
- Note that extensive gene overlap is not expected, and NCBI has several checks of annotation that may reject improperly annotated mitogenomes as well as properly annotated mitogenomes that diverge from exemplar reference sequences. If changes are not made after this annotation flag, NCBI will label the mitogenome as UNVERIFIED and will not appear in any BLAST matching results, dramatically decreasing their value as reference sequences.

- It is good practice to confirm the identity of your mitogenome assembly by blasting different genes to confirm the expected species identity of close matches.

- New tools are being developed to help assist in mitogenome assemblies, annotations, quality control and mediating the input of sequences into NCBI (MitoPilot, mtGrasp).

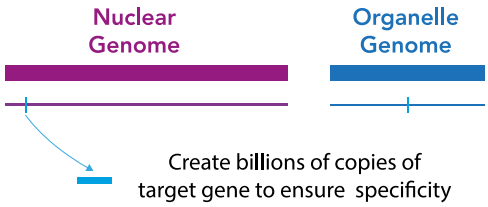
9. Comparison of Sanger Sequencing and Genome Skimming

In this section we provide an overview comparing sanger sequencing and genome skimming methods (Figure 2). We also present a handful of example situations, recommendations, and reasoning for why researchers and practitioners would choose sanger or genome skimming reference barcoding methods (Table 4) or whole genome sequencing approaches (Table 5). We also provide relevant information on the effort required to conduct sanger sequencing and genome skimming (Tables 6 and 7)

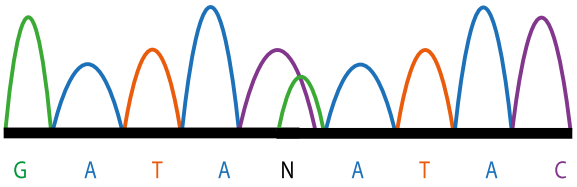


Sanger Sequencing

1. Amplify & Sequence Target Marker Gene



2. Quality Control of Resulting Chromatogram



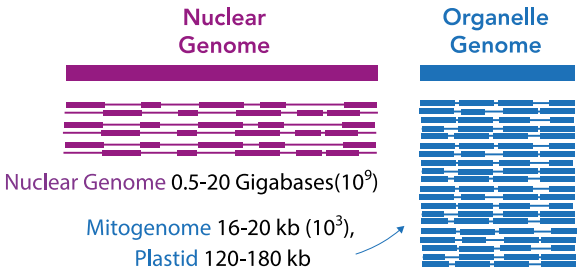
400-2000 bases of data

3. Validate Consensus Sequences

	Pros	Cons
Sanger Sequencing	Simpler & cheaper	Less comprehensive genetic resource
Mitogenome Skimming	More comprehensive genetic resource	Higher cost & complexity

Genome Skimming

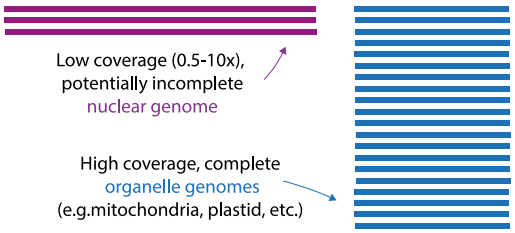
1. Low Coverage Shotgun Sequencing



2. Quality Control of Sequencing Data

5 Gigabases to 16 Terabases (10^{12}) of data

3. Assemble DNA Sequences



4. Annotate Sequences

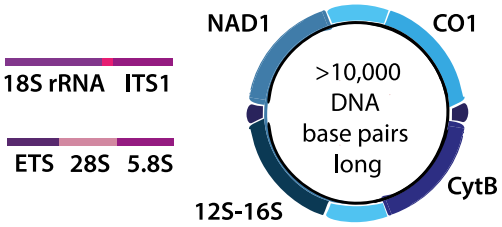


Figure 2. Comparison of Sanger Sequencing and Genome Skimming.

Recommendations

Situation	Recommendation	Reasoning
Specimen taxonomy is unclear based on morphology (e.g., Cryptic diversity within the clade), and marker gene assay is known to work	Sanger sequencing of marker gene	Important to identify species before selecting a representative for mitogenome skimming
No bioinformatics expertise or limited computational resources	Sanger sequencing of marker gene	Single marker gene analysis can be done with minimal prior training on a personal computer while bioinformatics expertise and moderate computing power is required for mitogenome assembly and annotation
Specific marker is identified as a priority (e.g., eDNA monitoring effort is already being employed)	Sanger sequencing of marker gene	Cost and time effective to focus on specific marker gene of interest
Specimen can be identified to the species level by morphology	Genome skimming	Basic taxonomy is already known; Genome skimming provides more robust data on multiple marker genes and better resources for subsequent work (e.g., eDNA)
Marker gene sequences for closely related species do not adequately resolve taxonomy	Genome skimming	Genome skimming provides more robust data needed for accurate taxonomic assignment [We note that for some species mitogenomes can not resolve taxonomy and additional nuclear genes will be needed]
Lowest cost : data ratio is desired (See Table 4)	Genome skimming	Genome skimming provides more robust data and offers resources for subsequent work well beyond taxonomy [Ever-decreasing sequencing costs make genome skimming more economically viable]
Data needed to design eDNA assays for species of interest (e.g., qPCR, dPCR, LAMP)	Genome skimming	Genome skimming provides data for multiple genes, increasing the likelihood of designing a successful eDNA assay

Table 4. A guide to inform decision-making for sequencing voucher specimens for reference barcoding.

Recommendations

Situation	Recommendation	Reasoning
Project goals include generating a robust phylogeny to map evolutionary histories	Whole genome sequencing	Single-gene phylogenies are often of limited value; concatenated marker genes and full mitogenomes provide better resolution, although many species' mitogenomes cannot resolve subtle phylogenetic relationships; full nuclear genomic data captures thousands of DNA variants needed to accurately resolve phylogenetic relationships
Project goals include robust population connectivity, stock assignment, or individual identification	Whole genome sequencing	Accurate assignment of individuals and stocks frequently requires tens to millions of DNA variants which are best obtained with whole genome sequencing data [We note whole genome sequencing data requires at least an order of magnitude more sequencing depth than genome skimming, increasing costs]

Table 5. A guide to inform decision-making for sequencing voucher specimens for population genetics.

At right: Giant kelp (*Macrocystis pyrifera*) is one of the target species for the California Conservation Genomics Program for whole genome sequencing. © Zack Gold

Comparison of Effort

Item	Sanger Sequencing	Genome Skimming	Whole Genome Sequencing
Extraction	$15-$35	$15-$35	$15-$35
Library Preparation + Sequencing	$5-$17	$120-$175	~$500-5,000

Table 6. Approximate cost comparison for Sanger sequencing a single marker gene vs genome skimming vs. whole genome sequencing. Cost estimates are derived from private sequencing companies in the U.S. as of August 2024 and are inclusive of labor and reagents. We note that whole genome sequencing has a high variability depending on the size of an organism's genome.

Labor	Sanger Sequencing	Genome Skimming
Manual extraction	~ 24 per 5-18 hours	~ 24 per 5-18 hours
Extraction by liquid handling robot	~ 96 per 5-18 hours	~ 96 per 5-18 hours
PCR + Gel	~ 4 hours	NA
Sanger Sequencing	~ 3 hours	NA
Library Preparation	~ 15 minutes to 1 hour	~ 3 hours
Genome Skimming	NA	~ 1-3 days on instrument
Bioinformatics	Minutes per individual	Hours to days per individual depending on computational resources and biological attributes
Submission to NCBI	Minutes per individual	Hours to days per individual depending on biological attributes, e.g., non-standard mitogenome architecture or no close reference sequences

Table 7. Approximate time comparison for labor required to generate Sanger sequences of single marker genes and genome skimming.

10. Archiving, Data Integration and Data Submission to GenBank

Prepare Data

Format sequence data and metadata according to NCBI GenBank's submission guidelines. Ensure all required fields are accurately completed. We strongly encourage adopting both Minimum Information about any (x) Sequence (MIxS) and DwC standards. Efforts are underway to align NCBI MIxS and Darwin Core to be more interoperable. See Resources Sections 3 and 4 for more information.

In all cases, generating a BioProject is a good idea so that multiple (potential future) sequences can be added and linked to the same organism/project.

For single genes, submission is limited to the final gene sequence.

For mitogenomes:

- Raw sequencing data should be submitted to NCBI SRA. Every SRA accession is associated with an NCBI BioSample record.
- Annotated mitogenomes should be submitted to NCBI GenBank with references to SRA and BioSample Accession numbers.
- Any other annotated reference from the organism assembly (i.e. rRNA-ITS nuclear genome region) should also be submitted to NCBI GenBank, with references to SRA and BioSample Accession numbers.
- For mitogenomes that cannot be completed or circularized, individual genes/gene regions that can be assembled can be uploaded individually, again with references to SRA and BioSample Accession numbers.

Submit Data

Upload raw sequence data to the **Sequence Read Archive (SRA)**. Submit other sequences (Sanger and annotated mitogenome) to GenBank following their submission procedures. There are linkouts from their **Submission Portal** to the different submission tools. For the sequences discussed here, that portal leads to a **BankIt** submission.

Make sure that all sequence submissions for a single organism lead to a single BioSample accession, and that all BioSample accessions for the project are linked to a single BioProject accession. This will simplify sharing and findability.

It may be necessary to email **gb-sub@ncbi.nlm.nih** for certain sequences or ask for information. Please don't hesitate to email these staff, they are very helpful, particularly with batch submissions.

table2asn is a command-line tool that creates sequence records for submission to GenBank.

If you submit a scientific name that is not already represented on GenBank, you can indicate that it is a known taxon. In this instance, we strongly encourage including a link to the appropriate taxonomic authority (e.g., World Registry of Marine Species, Catalogue of Life, or Global Names Architecture). If this species is new to science, no documentation is required at this stage; a placeholder taxon will be assigned and you can contact GenBank to update this when the new name is published.

Verify Submission

Confirm that your data has been successfully deposited with accurate metadata and the appropriate sequence release date (i.e. embargo). Even immediate release will take some time to become public on GenBank (particularly if a new taxID has to be created). On manuscript submission, make sure that the sequence data has been made public and that it is accessible through GenBank linkouts to the BioProject accession provided in your manuscript.

Archiving Specimens, Genomic DNA extracts, and Associated Data

Deposit physical specimens, genomic DNA extract, and associated data (compatible with Darwin Core standards and according to collection-specific requirements) with permanent collections. Ensure specimens are stored in archival-quality conditions. We advise doing this prior to any GenBank or International Nucleotide Sequence Database Collaboration (INSDC) submission to ensure discoverability and so that the holding instituional identifiers can be associated with the BioSamples and various submissions.

Many institutions and museums (e.g., **SIO Benthic Invertebrate Collection**, **SIO Pelagic Invertebrate Collection**, **Natural History Museum of Los Angeles County**, **the Smithsonian's** National Museum of Natural History **Invertebate Zoology Collection**) publish digital catalogs of their holdings through institutional websites and biodiversity aggregators such as iDigBio, OBIS, and GBIF, thereby increasing the discoverability and utility of the specimens and data.

Link and Integrate Datastreams

Link your sequence and specimen data to related datasets and maintain consistency across databases. Ensure that all records are integrated and accessible in publicly available repositories.

Taxonomists identifying organisms and maintaining data records to link identifications to specimens and collection events. © NHMLAC

Additional Considerations

Quality Control

Regularly check and verify data accuracy through manual reviews and reanalysis if necessary. If updates are required for information already on NCBI, contact their staff for update procedures that are generally straightforward.

Ensure that all procedures adhere to best practices and standardized protocols.

Data Integration

Utilize tools and standards such as Darwin Core for data linkage and interoperability, making sure data is Findable, Accessible, Interoperable, and Reusable (FAIR).

Documentation

- Maintain comprehensive records at each step of the process to support data integrity and reproducibility.
- Rigorous and detailed data and metadata documentation are needed to make the data as valuable as possible.
- Include details from collection through archiving to ensure complete and useful data sets.

Development of Standard Reference Materials

Vouchers are important for developing biological standard reference materials, particularly mixed composition samples (e.g., mock communities) and for samples of unknown provenance (i.e. artifacts and processed tissues/parts). Standard reference materials require the greatest level of trust and accuracy and thus careful consideration and quality assurance is needed to develop such resources.

We also acknowledge that some mock communities and standard reference materials may be used for next-generation sequencing approaches. The high sensitivity of these approaches poses a greater risk for contamination. We caution researchers to ensure such standard reference materials are generated and processed in a way that ensures their long term viability and widespread applicability.

Protocols and Best Practices

Standardization

- **Protocols:** Use standardized protocols across institutions for consistency in specimen handling, preservation, and data entry.
- **Data Templates:** Implement clear templates for recording and managing specimen and sequence data. Adopt existing systems where feasible, such as **GEOME**.

Quality Control

- **Manual Checking:** Conduct manual reviews of sequence data to identify and correct errors. Verify that sequences align with reference data and are free of contaminants.
- **Data Integrity:** Ensure that all data elements are correctly linked and that the hierarchy is maintained.
- **Reanalysis:** Use industrialized rerun protocols if necessary to re-evaluate data within comprehensive data hierarchies.

Resources

1. Open Science Standards

1.1 Global Indigenous Data Alliance - **CARE** Data Standards

1.2 Go **FAIR** - FAIR data Standards

1.3 NOAA Scientific Data Stewardship Public Access to Research Results (**PARR**)

2. Comprehensive Materials

2.1 Neumann, D., Carter, J., Simmons, J. E., and Crimmen, O. (2022). **Best practices in the preservation and management of fluid-preserved biological collections**.

2.2 Global Genome Initiative Expeditionary Collecting **Training Module**

2.3 Cold Spring Harbor Laboratory DNA Learning Center **DNA Barcoding 101**

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

2.5 **Curation Procedures Applicable to BOEM Invertebrate Collections Transferred to the Smithsonian Institution's National Museum of Natural History**

2.6 Smithsonian Institution's National Museum of Natural History and Bureau of Ocean Energy Management – Environmental Studies Program (BOEM-ESP) **Genetic Resources and Documentation**

2.7 DeSalle, Robert, editor. **DNA Barcoding: Methods and Protocols**. 1st ed., Springer US, 2024.

2.8 Abdi, G. et al. (2024). **DNA Barcoding and its Applications**. In: Singh, V. (eds) Advances in Genomics. Springer, Singapore.

3. Darwin Core Standards:

3.1 Wieczorek et al. 2012. **Darwin Core: An Evolving Community-Developed Biodiversity Data Standard**

3.2 Horton et al., 2021 **Recommendations for the Standardisation of Open Taxonomic Nomenclature for Image-Based Identifications**.

3.3 Meyer et al. 2021. **Aligning Standards Communities: Sustainable Darwin Core MIxS Interoperability**

3.4 NOAA IOOS **Marine Biodiversity Data Workshop**

3.5 OBIS Manual **Darwin Core Terms and Guidelines**

3.6 GBIF **What is Darwin Core, and why does it matter?**

3.7 USGS **Introduction to Darwin Core**

4. Relevant Biodiversity, Molecular, and Ocean Observing Standards

4.1 Taxonomic Databases Working Group's (**TDWG**) **Darwin Core**

4.2 Genomic Standards Consortium's (**GSC**) Minimum Information about any (x) Sequence (**MIxS**) and The minimum information about a genome sequence (**MIGS**) specification

4.3 Global Ocean Observing System's (**GOOS**) **Essential Ocean Variables**

4.4 Global Earth Observing System of Systems (**GEOSS**) Standards

4.5 National Microbiome Data Collaborative's (**NMDB**) Standards for Technical Reporting in Environmental and host-Associated Microbiome Studies (**STREAMS**).

5. Data Management

5.1 Parker et al. 2012. **Laboratory Information Management Systems for DNA Barcoding**

5.2 USGS **Guide to Planning for and Managing Scientific Working Collections**

5.3 Trizna 2018. **Best practices for connecting genetic records with specimen data**

5.4 Rimet et al. 2021. **Metadata Standards and Practical Guidelines for Specimen and DNA Curation**

5.5 **Extended Specimen Network**

5.6 FIMS, LIMS and Geneious Integration: Utilize **Laboratory Information Management Systems (LIMS)** for efficient data management and integration. The Smithsonian's National Museum of Natural History uses the freely available **Biocode LIMS Plugin** that works with Geneious software in association with the freely available **GEOME** Field Information Management System (FIMS) that manages event, specimen and tissue data and metadata following DwC standards.

6. Specimen Collection

6.1 ABC Taxa, Volume 8 - **Manual on Field Recording Techniques and Protocols for All Taxa Biodiversity Inventories**

6.2 Rouse, G., Pleijel, F., and Tilic, E. (2022). **Annelida**.

6.3 Martin JW 2016. **Collecting and Processing Crustaceans: An Introduction**

6.4 Bentley et al. 2024. **Community Action: Planning for Specimen Management in Funding Proposals**

6.5 Motomura and Ishikawa 2013. **Collection Building and Procedures Manual, English Edition**

6.6 Buckner et al. 2021. **The critical importance of vouchers in genomics**

6.7 Culley 2013. **Why Vouchers Matter In Botanical Research**

6.8 **CA Department of Fish and Wildlife Scientific Collecting Permits**

6.9 International Air Transport Association (IATA) **Dangerous Goods**

6.10 Natural History Museum of Los Angeles County Diversity Initiative for the Southern California Ocean (DISCO) **Marine Invertebrate Collecting Protocols**

7. Laboratory and Sequencing

7.1 Stoekle and Herbert 2008. **Barcode of Life**.

7.2 Theissinger et al. 2023. **How genomics can help biodiversity conservation**.

7.3 Antil et al. 2023. **DNA barcoding, an effective tool for species identification: a review**.

7.4 Hoban et al. 2022. **Skimming for barcodes: rapid production of mitochondrial genome and nuclear ribosomal repeat reference markers through shallow shotgun sequencing**

7.5 Quattrini et al. 2024. **Skimming genomes for systematics and DNA barcodes of corals**

7.6 Trevisan et al. 2023. **Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies**

7.7 Levesque-Beaudin et al. 2023. **A workflow for expanding DNA barcode reference libraries through 'museum harvesting' of natural history collections**.

7.8 Bemis et al. 2023. **Biodiversity of Philippine marine fishes: A DNA barcode reference library based on voucher specimens**.

7.9 Gold et al. 2021. **Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem**.

7.10 McLaughlin et al. 2021. **Resolving the taxonomy of the Antarctic feather star species complex Promachocrinus 'kerguelensis' (Echinodermata: Crinoidea)**.

7.11 National Human Genome Research Institute 2022. **DNA Sequencing Costs: Data**

7.12 Millipore Sigma **Sanger Sequencing Costs and Method**

8. Bioinformatics Software

8.1 **SPAdes**

8.2 **getOrganelle**

8.3 **phyloFlash**

8.4 **mtGrasp**

8.5 **PMAT**

8.6 **Minimap2**

8.7 **Magic-BLAST**

8.8 **MITOS2**

8.9 **MitoZ**

8.10 **table2asn**

8.11 **MitoPilot**

9. Archiving Genetic Sequences

9.1 NCBI Sequence Read Archive (**SRA**)

9.2 NCBI GenBank **Submission Portal**

9.3 NCBI **BankIt** submission portal

9.4 European Nucleotide Archive (**ENA**)

9.5 Barcode of Life Data Systems (**BOLD**)

10. Relevant U.S. National Strategies

10.1 **National Aquatic eDNA Strategy**

10.2 **National Ocean Biodiversity Strategy**

10.3 **The US Ocean Biocode**

11. Collaborations

11.1 **West Coast OBON** a project within the **OBON Program** within the UN Decade of Ocean Science for Sustainable Development. We work toward method harmonization and data integration of 'Omics ocean observing platforms on the North American West Coast in support of sustainable marine management. Efforts include establishing best practices for large scale biomolecular monitoring and sequencing mitogenomes of ecologically important taxa to improve DNA reference libraries.

11.2 **California Conservation Genomics Program** brings together many of California's leading experts working at the interface of genomics and conservation science. They are a state-funded initiative with a single goal: To produce the most comprehensive multispecies genomic dataset ever assembled to help manage and protect regional biodiversity in the face of climate change.

11.3 **MetaZooGene** a SCOR working group for generating reference barcodes for zooplankton and their application to molecular based identification of zooplankton species.

11.4 **BOEM and Smithsonian Partnership**. The Smithsonian's world-class analysis of BOEM's and other offshore invertebrate collections increases understanding of marine biodiversity and environmental change. BOEM and the Smithsonian Institution have maintained a long and productive partnership since 1979, greatly benefitting both agencies, the scientific community worldwide, and, by extension, the public's access to scientific information.

11.5 **The National Systematics Laboratory of NOAA Fisheries and Smithsonian NMNH**. Embedded within the Smithsonian NMNH, this partnership, formally established in 1942, supports experts on marine taxa and collections working closely with NMNH counterparts to advance systematic, taxonomic and life history research on marine organisms of ecological and economic value to the United States. This partnership furthers both agency missions and results in vital knowledge on ocean biodiversity being put into the public domain.

11.6 **NMNH Ocean DNA Project**: Ocean DNA is a museum-wide effort to leverage national collections, partnerships and adopt museum practices to support the use of DNA sequencing to survey marine life and assess ocean health, with a special emphasis on undiscovered "dark taxa" that may play a critical role in ecosystems.

DOI # 10.5281/zenodo.14867763

https://evsatt.github.io/WC-OBON_Website/