E&P **ENVIRONMENTAL GENOMICS RESEARCH** PROGRAMME

An IOGP programme

International Association of Oil & Gas Producers

DRAFT | November 2024

Industry Guidance on Bioinformatics Analysis Standards and Guidelines for eDNA Data relevant to O&G

This Guidance is publicly available as a draft version for readers for review and trial its use. Any comments that improve the content or workability of the draft guidance are welcomed by the JIP.

Please send comments to **info@iogp-edna.org**.

We will endeavour to address comments, if received by Feb 28th 2025, in the Final (edited and formatted) version of the guidance which will be available early in 2025. Any comments received after this date or any comments that cannot be addressed within this timeframe will be considered for future updates to the guidance. Updates are planned every 1-2 years to capture the rapid advances in this field.

Prepared by CEGA and eDNAtec for the International Oil and Gas Producers' Association Joint Industry Program on Environmental Genomics.

REVISION HISTORY

RELEASE	NAME	COMPANY	DATE OF ISSUE	COMMENTS
R1	Greg Singer	eDNAtec	2023-Sep-04	First (draft) release.
R2	Greg Singer	eDNAtec	2023-Nov-24	Initial round of revisions to address IOGP reviewer comments.
R3	Greg Singer	eDNAtec	2024-Jan-28	Revisions throughout to address additional IOGP reviewer comments.
R4	Greg Singer	eDNAtec	2024-Jul-09	Revisions throughout to address further reviewer comments and integrate text from RFP2. Introduce executive summary and Introduction.
R5	Nicolas Tsesmetzis	Shell	2024-Oct-11	Addition of Revision History and Disclaimer
R6	Michael Marnane	Chevron	2024-Nov-21	Addition of Instructions to readers

Table of Contents

Table of abbreviations and definitions

Abbreviation	Definition
Amplicon	A piece of DNA that is the product of a polymerase chain reaction. Within the context of metabarcoding, a segment of a gene that will be used for DNA barcoding purposes
ASV	Amplicon Sequence Variant – a unique nucleic acid sequence obtained from the sample
BLAST	Basic Local Alignment Search Tool – a very popular algorithm to find targets in a sequence database that are similar to a query, based on sequence similarity measures
Chimeras / Chimeric sequences	False sequences formed by the incorrect joining of two or more biological sequences together. This often occurs during PCR.
COI	Cytochrome c Oxidase subunit I, a gene within the mitochondrial genome that is a popular marker for DNA barcoding studiesI
CPU	Central Processing Unit, the chip in each computer that is primarily responsible for general computational tasks
ddPCR	Droplet Digital PCR, a method of PCR that allows the direct quantification of targets within a sample
DNA	Deoxyribose Nucleic Acid, the primary information storage molecule in cells
eDNA / eRNA	Environmental DNA (or RNA), which is derived isolated from environmental samples
GBIF	Global Biodiversity Information Facility – a free database that contains hundreds of millions of species occurrence records
GPU	Graphical Processing Unit, which is faster for certain types of computation than the CPU
MAG	Metagenome-assembled genomes – genomes that have been reconstructed from metagenomic data
MOTU	See OTU
NMDS	Non-parametric Multidimensional Scaling -— a method for highlighting gradient structures within high-dimensional data. Also see PCoA

ONT	Oxford Nanopore Technologies, makers of the popular Nanopore sequencing platform
OTU / MOTU	Operational Taxonomic Unit – groups of closely related sequences
PCoA	Principal Coordinate Analysis – a statistical method for highlighting structures within high-dimensional data. Also see NMDS.
PCR	Polymerase Chain Reaction – a molecular biology technique for replicating DNA sequences
QC	Quality Control, a series of tests to ensure results are adhering to the expected standards
qPCR	Quantitative PCR, sometimes also called real-time PCR (RT-PCR) which should not be confused with reverse-transcriptionase PCR (RT-PCR), the method of converting RNA sequences into DNA sequences. qPCR is used to quantify the amount of a target sequence within a sample (also see ddPCR)
RAM	Random Access Memory – a measure of volatile memory within a computer
RNA	Ribose Nucleic Acid, a short-lived molecule that is derived from nuclear DNA and typically encodes gene sequences

Executive Summary

This guideline provides a comprehensive overview of the role of bioinformatics in environmental genomics analyses. It outlines the standardized workflow involved in analyzing metabarcoding data, including demultiplexing, quality filtering, denoising, read merging, artifact filtering, clustering, and taxonomic assignment.

Quality control (QC) is crucial in metagenomics and metabarcoding workflows to ensure reliable and actionable results. High-quality sequence data impacts taxonomic assignments and computational efficiency. Effective QC minimizes the risk of spurious detections and supports reproducibility across datasets. Various QC tools tailored to different sequencing platforms (e.g., Illumina, PacBio, Oxford Nanopore), each with their own strengths and limitations in optimizing sequence data quality for robust scientific analysis.

Denoising is an important step to reduce impact of single nucleotide variants (SNVs) on taxonomic assignments. SNVs, often artifacts from PCR or sequencing processes, can lead to inaccuracies in targeted sequencing analyses. Methods like operational taxonomic units (OTUs) and amplicon sequence variants (ASVs) mitigate these errors by clustering or denoising sequences, respectively. OTUs group similar sequences to minimize errors and dependency on reference databases, while ASVs preserve individual sequence uniqueness and are more reproducible.

Taxonomic assignment is one of the more challenging steps in the analytical pipeline. Confounding issues include PCR and sequencing errors, incomplete reference databases, and the need for accurate algorithm selection. Algorithms are crucial for probabilistically determining taxonomic assignments based on sequence similarity. Methods like k-mer based approaches (e.g., Kraken2) and sequence similarity tools (e.g., MegaBLAST) offer rapid taxonomic assignment but may sacrifice accuracy for speed. Phylogenetic approaches (e.g., EPA-NG) provide the highest accuracy by placing query sequences within evolutionary contexts but are computationally intensive. A nuanced approach to algorithm selection is recommended, based on specific project goals and dataset characteristics. Faster algorithms can be used for initial taxonomic assignments, while phylogenetic methods should be employed for resolving ambiguous results or when accuracy is paramount. Integrating multiple markers and validation through consensus methods can enhance taxonomic reliability, especially for invasive or endangered species detection.

Reference database choice significantly impacts the assignment of taxonomy to sequence reads and the quality thereof. Curated databases ensure higher accuracy but often limit the number of taxonomic assignments due to stringent curation standards. Uncurated databases like GenBank's nucleotide database provide broader taxonomic coverage but at the expense of accuracy. The decision on reference database choice should align with study objectives, leveraging comprehensive databases like GenBank for broader biodiversity assessments versus curated databases for precise species identification.

Managing false positive and false negative detections in metabarcoding data is crucial for accurate biodiversity assessments. False positives can arise from contamination, while false negatives can result from PCR inhibition or insufficient sequencing depth. Balancing false positives and negatives is essential to maintain a degree of sensitivity required to meet study objectives. Hierarchical occupancy modeling helps correct for detection errors, supporting robust conclusions.

Quantitative analyses in biodiversity studies are challenging with environmental DNA (eDNA) because it is subject to various biases. Laboratory processes and environmental factors can introduce biases in sequence read counts, complicating the accurate interpretation of data. Despite these challenges, strong correlations between sequence read counts and organism biomass or abundance are frequently demonstrated. Techniques like occupancy modeling and normalization methods (e.g., rarefaction curves) help manage variation in sampling effort and ensure robust biodiversity estimates.

Bioindicators and biotic indices are essential for monitoring ecosystem health. Environmental genomics aids in identifying both known and novel bioindicator taxa, capturing comprehensive community data. Biotic indices consolidate bioindicator information into single metrics, crucial for assessing ecosystem quality. Environmental genomics data support existing biotic indices and facilitate the development of new ones, yet interpretation requires awareness of differences from traditional morpho-taxonomic data.

Understanding the functional characteristics of an ecosystem is possible through metagenomics and metatranscriptomics. These techniques are most highly developed for microbial community analysis, although the techniques are increasingly being applied to metazoan communities.

The field of environmental genomics is rapidly progressing, driven by expanded project scopes, increased sequencing depths, and growing reference databases. Advanced algorithmic and computational solutions are enhancing data analysis speeds. However, gaps in reference databases persist, requiring focused efforts to bridge them. Establishing standardized reporting for eDNA data is crucial for enhancing confidence in results and supporting broader adoption in industry and regulatory contexts.

Decision tree for navigating this document

Step 1: Base calling and demultiplexing

Decision 1: What sequencing platform is being used

- Want a portable solution and long read-lengths: Oxford Nanopore
    - Speed is more important than accuracy → use Guppy (Section 2.3.1.1)
    - Accuracy is more important than speed → use Bonito (Section 2.3.1.2)
- Want to sequence a large number of reads from the sample: Illumina
    - Use BCL-Convert (Section 2.3.2.2)
- Want long read-lengths: PacBio
    - Use SMRT Analysis (Section 2.3.3.1)

Decision 2: What type of study is this

- Want to understand the biological function characteristics of the environment: Metagenomics / Metatranscriptomics → go to Section 5
- Want to know what species are present in the environment: Metabarcoding → go to Step 2

Step 2: Quality control

Decision 1: What sequencing platform is being used (from Step 1)

- Oxford Nanopore → see Section 2.4.2.3
- Illumina → see Section 2.4.2.1
- PacBio → see Section 2.4.2.2

Step 3: Clustering/denoising

Note: clustering/denoising software will typically also merge paired-end reads (if it is applicable to your experiment). Merging of pairs is sometimes recommended before clustering/denoising (e.g., VSEARCH) or after (e.g., DADA2).

Decision 1: What type of error reduction is desired

- Increased data reduction, increased processing speed, decreased precision in taxonomic assignments → OTU clustering (Section 2.5.2)
- Less data reduction, slower processing speed, increased precision in taxonomic assignments → denoising (Section 2.5.5)

Step 4: Taxonomic assignment

Decision 1: Choice of reference database

- Matching the highest proportion of reads is paramount (e.g., general biodiversity survey) → use an uncurated database (see Section 3.3)
- Highest accuracy is paramount (e.g., looking for specific organisms)
    - An existing reference database exists that suits the needs of the project → use a curated reference database (see Section 3.4)

    - A new reference database will need to be generated → make a custom reference database (Section 3.5)

Decision 2: Choice of taxonomic assignment algorithm

- Speed is most important → use a k-mer algorithm (see Section 2.7.2)
- Balance of speed and accuracy → use one of the BLAST algorithms (Section 2.7.3)
- Accuracy is most important (access to high-performance computing resources is available) → use a phylogenetic method (Section 2.7.4)

Step 5: Data analysis/interpretation

Decision 1: are quantitative results desired → see Section 4.1.3

Decision 2: if a biotic index will be used → Section 4.2.2

Decision 3: if you need to understand differences between biological communities

- The dataset is small in size → Section 4.2.3.2
- The dataset is large and:
    - Want to understand the interactions between species, their traits, and their phylogenetic relationships → Section 4.2.3.3.1
    - A high confidence in detection/non-detection is needed → Section 4.2.3.3.2
    - Want to perform a network analysis on species relationships → Section 4.2.3.4

1 Introduction

Bioinformatics plays a critical role in environmental DNA (eDNA) analysis, which involves the extraction and examination of genetic material from environmental samples such as soil, water, or air to study biodiversity. The process starts with the raw DNA sequence data that has been generated by laboratory analyses. The process encompasses several key steps:

1.1 Data Processing and Management

- **Sequencing Data Handling:** High-throughput sequencing generates large volumes of data. Bioinformatics tools are used to efficiently manage, store, and preprocess these data. This includes tasks such as demultiplexing, which sorts sequences from different samples, and removing adapter sequences.
- **Quality Control (QC):** Ensuring high-quality data is fundamental. Tools like FastQC, Trimmomatic, and Cutadapt are employed to filter out low-quality reads, trim adapters, and remove contaminants. QC is crucial for accurate downstream analysis.

1.2 Sequence Analysis

- **Denoising and Error Correction:** Bioinformatics algorithms correct errors introduced during PCR amplification and sequencing. Tools like DADA2 help distinguish true biological sequences from artifacts.
- **Read Merging:** Paired-end reads from sequencing can be merged to form longer contiguous sequences, enhancing the accuracy of subsequent analyses.

1.3 Taxonomic Assignment

- **Clustering and OTU/ASV Formation:** Sequences are grouped into Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs). OTUs cluster sequences based on similarity, while ASVs represent individual sequences after denoising.
- **Taxonomic Classification:** Bioinformatics tools assign sequences to taxonomic groups by comparing them against reference databases. Methods vary from k-mer-based approaches (e.g., Kraken2) to alignment-based methods (e.g., BLAST) and phylogenetic approaches (e.g., EPA-NG). Each method has trade-offs in terms of speed, accuracy, and computational requirements.

1.4 Functional Annotation

- **Functional Profiling:** Beyond identifying species, bioinformatics allows for the prediction of functional potential within a community. Tools like PICRUSt and Tax4Fun infer the presence of genes and metabolic pathways from taxonomic data, while direct functional annotation can be performed using tools like HUMAnN2 and eggNOG.

1.5 Data Integration and Interpretation

- **Statistical Analysis:** Bioinformatics provides statistical frameworks to analyze the diversity and structure of microbial communities. Techniques include alpha and beta diversity metrics, multivariate analyses, and differential abundance testing.

- **Modeling and Visualization:** Bioinformatics tools assist in visualizing complex data through heatmaps, ordination plots, and network analyses.

Bioinformatics is indispensable in eDNA analysis, driving the entire process from raw data to meaningful ecological insights.

2 Metabarcoding pipeline

2.1 Introduction

Currently, metabarcoding is overwhelmingly the most common approach to obtaining comprehensive biodiversity data from environmental DNA. Although there are different paths to take through the workflow and some steps are optional, the overall series of steps is generally quite consistent. Figure 1 is from a recent review paper about various pre-built pipelines available (Hakimzadeh et al., 2023), and illustrates the general workflow and the alternative paths that can be taken. Here, we give a brief overview of each step and in the next few sections, we proceed through the workflow step-by-step and consider the various options available and make recommendations about when one option may be the superior choice in a given situation.

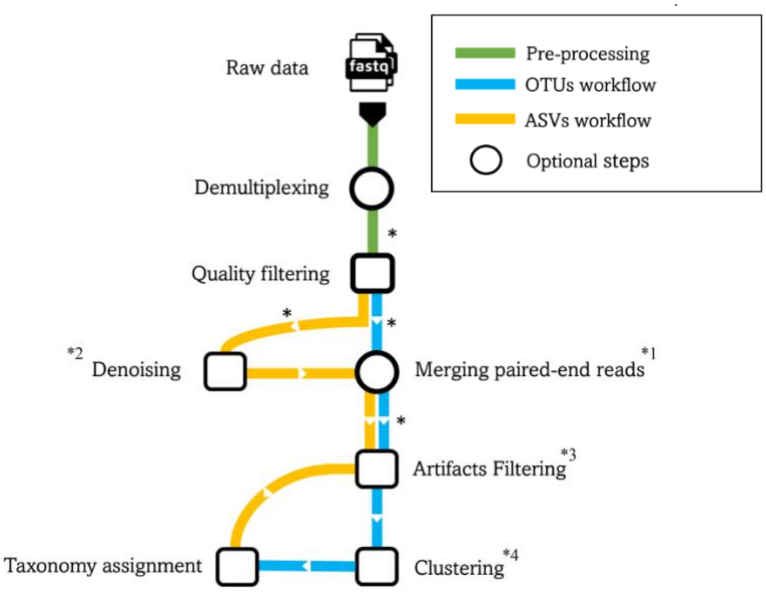


Figure 1: The typical metabarcoding analytical pipeline workflow, with a few alternative paths and optional steps (Hakimzadeh et al., 2023)

For a given metabarcoding project, it is important that all samples are processed with the exact same bioinformatics pipeline, as different bioinformatics pipelines and parameters can potentially give significantly different results from the same raw sequence data. It is also particularly important to consider the need to link together datasets generated from different sequencing runs. This is especially relevant for taxonomic groups and markers with incomplete reference databases, meaning taxa cannot be linked based on species names and may influence the choice between use of OTUs (operational taxonomic units) and ASVs (Amplicon Sequence Variants; (Callahan et al., 2016)), discussed later in this document (see Section 2.5.1).

2.1.1 Steps in the metabarcoding pipeline (refer to Figure 1)

- Demultiplexing – a procedure that sorts DNA sequences that have been processed together into individual samples/projects
- Quality filtering – a series of steps to remove bad DNA sequence reads (e.g., with poor base calling scores, inappropriate lengths, etc.)
- Denoising – a statistical technique that attempts to remove errors introduced through the polymerase chain reaction or sequencing laboratory steps
- Merging paired-end reads – in cases where sequencing is bi-directional and there is an overlap between the forward and reverse reads, these can be merged into a single sequence
- Artifacts filtering – an additional QC step to look for chimeras (i.e., the forward sequence from one amplicon being merged with the reverse sequence from another)
- Clustering – grouping similar DNA sequences together
- Taxonomic assignment – attempting to identify the organism from which each DNA sequence came from

2.2 Comparisons of pre-existing pipelines

Pipelines are the heart of most bioinformatics analyses. In a broad sense, a pipeline can be described as the "glue" connecting a series of software packages to perform an analysis. Some pipelines can be end-to-end, starting at demultiplexing and ending with taxonomic assignment and others cover more narrow workloads. As the implementation of a pipeline from scratch is a non-trivial task, a plethora of downloadable pipelines have been published, offering unique algorithms and methodologies tailored for different aspects of sequence processing. This section of the guidebook will cover and compare published pipelines that can be used for analyzing metabarcoding data. We will focus on end-to-end pipelines and evaluate these pipelines over a range of relevant criteria. Our analysis also includes discussions on the considerations for choosing a pipeline based on specific project goals, such as the taxonomic breadth of the study, the size and complexity of the dataset, and the available computational resources. By dissecting the strengths and limitations of each pipeline, we provide a resource that can inform researchers' pipeline choices, thereby optimizing the accuracy and efficiency of metabarcoding studies.

2.2.1 Pros and cons of using a pre-existing pipeline

The choice between utilizing a pre-existing pipeline package and developing a custom pipeline is a strategic decision that researchers must make, weighing various technical and practical considerations. Here we discuss the advantages and disadvantages of adopting pre-existing metabarcoding pipelines versus the construction of a bespoke pipeline.

2.2.1.1 Using a Pre-existing Pipeline

Pros:

1. **Time-Efficient**: Pre-existing pipelines are ready to use, which significantly reduces the time from sample collection to data analysis and interpretation.

2. **Community Validation**: These pipelines have often been tested and validated by a large community, which can add confidence in the reproducibility and reliability of results.

3. **Support and Documentation**: A broad user base ensures better support and extensive documentation, which can help troubleshoot and guide new users through the process.
4. **Regular Updates**: Maintained pipelines benefit from regular updates that include bug fixes, new features, and improvements to keep pace with the evolving field.
5. **Cost-Effective**: They can be more cost-effective since developing a new pipeline requires significant investment in terms of time and resources.

Cons:

1. **Generalization Over Specialization**: Pre-existing pipelines may not cater to specific, unique, or novel requirements of certain projects.
2. **Learning Curve**: There may still be a learning curve associated with using complex pipelines, which can be intimidating for new users. In addition, a user may need to learn a specific programming or workflow language they are not familiar with to interface with the pipeline.
3. **Inflexibility**: Some pipelines may lack flexibility, forcing users to adapt their research questions or data collection methods to the requirements of the pipeline.

2.2.1.2 Developing Your Own Pipeline

Pros:

1. **Customization**: Developing a pipeline from scratch allows researchers to tailor the tool to their specific needs, which can be crucial for novel or unconventional studies.
2. **Optimization**: A custom pipeline can be optimized for the specific data types, quality, and analysis workflows of a particular project. It may also be optimized on a hardware level for computational efficiency.
3. **Innovation**: Creating a new pipeline contributes to the field by providing solutions to unaddressed problems and sharing new methodologies with the community.

Cons:

1. **Time and Resource Intensive**: Designing and testing a new pipeline requires a significant time investment and substantial computational resources for testing and validation.
2. **Expertise Required**: A high level of expertise in bioinformatics, programming, and statistics is necessary to develop a functional and reliable pipeline.
3. **Lack of Initial Validation**: A newly developed pipeline lacks the extensive validation that established pipelines have, which may lead to skepticism or the need for extensive proof to gain acceptance during peer review and publishing.
4. **Maintenance and Support**: Developers must commit to maintaining the pipeline, fixing bugs, and providing user support, which can be a continuous and exhaustive task.

2.2.1.3 Balanced Consideration

Choosing whether to use a pre-existing pipeline or develop a new one is a decision that should be balanced against the goals, scale, and scope of the project. Pre-existing pipelines offer a plug-and-play solution with established support and community trust, suitable for standard and broad-scope projects. However, for cutting-edge research that pushes the boundaries of current knowledge, or where data types are novel or highly specialized, developing a new pipeline might be the best approach, despite the associated costs and efforts.

In some cases, a hybrid approach may be beneficial, where researchers start with a pre-existing pipeline and modify or extend it to meet their specific needs. This can combine the advantages of having a stable base to work from with the flexibility to innovate and customize as required.

Ultimately, the decision should consider the long-term benefits versus the immediate costs, with a clear understanding that the chosen approach aligns with the overarching research objectives and resource availability.

2.2.2 List of Pre-existing Pipelines

When searching the literature for published metabarcoding pipelines there will no shortage of options. A recent paper has reviewed 32 pipelines suitable for metabarcoding data analyses (Hakimzadeh et al., 2023). While tools such as DADA2 and USEARCH can be considered "pipelines" in their own right, as they are in the aforementioned paper, for this section we focus on pipelines that encapsulate tools such as these to produce a more end-to-end and user-friendly solution. In addition, we limit the discussion to pipelines that have been updated within the last 3 years and offer something unique or substantive to the user, to avoid comparing pipelines with only minor or superficial differences. Therefore, we limit this comparison to a handful of established options and offer an explanation on how these pipelines differentiate themselves and why they may be worthy of consideration.

QIIME 2

QIIME 2 (Quantitative Insights Into Microbial Ecology 2) is an open-source software package that provides a suite of tools for managing, analyzing, and visualizing microbiome data, primarily from DNA sequencing studies (Bolyen et al., 2019). It is an updated version of the original QIIME software package, which was widely used for analyzing microbial communities, usually obtained through sequencing the 16S rRNA gene, among other marker genes. Researchers may consider QIIME 2 if they want flexibility in which tools they use for an analysis as QIIME 2 offers multiple options for most workloads, such as offering both dada2 and deblur denoisers, as well as multiple options for taxonomic assignment.

Strengths:

1. **Comprehensive Analysis:** QIIME2 supports the entire analysis pipeline from raw data to the final visualization.
2. **Flexibility:** It offers a wide range of plugins and can be adapted to different analysis needs.
3. **Community Support:** A strong community of users and developers, which makes it easier to find help and resources.
4. **Interactive Visualizations:** Visualization tools can help with understanding complex data.

5. **Reproducibility:** Tracking of data provenance aids in reproducibility and transparency of research.

Weaknesses:

1. **Learning Curve:** QIIME2 has a steep learning curve for those new to bioinformatics or metabarcoding. In addition, it introduces new concepts such as the .qza format not present in other pipelines.
2. **Computational Resources:** Some analyses may require significant computational resources, which can be a limitation for some users.
3. **Updates and Changes:** Frequent updates and changes in the software may require users to continuously learn new interfaces or workflows.
4. **Plugin Quality:** While many plugins are available, the quality and maintenance can vary.
5. **Documentation:** Although extensive, documentation can be overwhelming and may not cover all specific use cases or issues.

Ampliseq

The Ampliseq software package is a bioinformatics pipeline for processing amplicon sequencing data (Straub et al., 2020). It supports denoising and taxonomic assignment for various gene regions such as 16S, ITS, CO1, and 18S, with phylogenetic placement capabilities. It's designed for both paired-end and single-end data from Illumina, PacBio, and IonTorrent sequencing technologies. Researchers may consider using Ampliseq if they are interested in using the Nextflow workflow management system, which offers a variety of benefits such as easy deployment on the cloud.

Strengths:

- Supports multiple sequencing platforms and amplicon types.
- Built with Nextflow, ensuring portability and reproducibility using Docker/Singularity containers.
- Automated quality control, read trimming, ASV inference, and taxonomic classification.
- Includes features for excluding unwanted taxa and differential abundance testing.
- Results are reproducible with defined resource allocations and persistent storage for benchmarking.

Weaknesses:

- A learning curve for setting up and running Nextflow and related container technologies.
- Potential for high computational resource requirements depending on the dataset size.
- Although QIIME 2 is wrapped, the full power and flexibility of QIIME 2 by itself is lost without substantial modification to the pipeline.

APSCALE

APSCALE is another pipeline that can be used for merging paired-end sequences, trimming primers, filtering by quality, clustering, and denoising (Buchner et al., 2022). It can handle popular metabarcoding markers such ITS, 16S ribosomal RNA, and COI. The software differentiates itself by having a focus on accessibility, with GUI (graphical user interface) for biologists without extensive bioinformatics expertise, facilitating reliable analysis of metabarcoding data for biodiversity research and environmental management.

Strengths:

1. **Platform Independence:** Can be run on Linux, macOS, and Windows.
2. **User-Friendly:** Offers both a command-line and a GUI version, with a simple installation process via Python's package installer, pip.
3. **Scalability:** Designed to handle large datasets efficiently, with support for multithreading.
4. **Data Protection Compliance:** Runs locally, catering to environmental agencies' data privacy needs.
5. **Comprehensive Data Processing:** Provides a wide range of functions from paired-end merging to taxonomic assignment.
6. **Visualization and Summary Statistics:** Enables visualization and statistical analysis within the GUI.

Weaknesses:

1. **Dependency on Python:** The GUI requires python to be installed and available which limits it's ease-of-use for biologists with no command line experience.
2. **Manual Demultiplexing:** Demultiplexing is not handled directly by APSCALE, requiring additional steps or software.
3. **Manual Installation of Dependencies for Windows:** Windows users must manually install the zlib library.
4. **Local Data Storage:** Large datasets may require significant local storage space as the GUI is not suitable for headless server use.

MetaWorks

MetaWorks is a bioinformatic pipeline designed for using a variety of metabarcoding markers such as 16S, ITS, COI, rbcL, 12S, and 18S (Porter & Hajibabaei, 2022). The tool operates in a Conda environment and is automated using Snakemake, which minimizes user intervention and facilitates scalability, making it suitable for use on high-performance computing clusters or the cloud. It utilizes the RDP Classifier for taxonomic assignments with confidence measures. In addition, it provides curated databases for the various supported markers available to download for this classification. Researchers may consider this pipeline if they have a wide variety of markers in their data and want to make use of the Snakemake

workflow management system. As Snakemake workflows are python code, they may be more familiar to bioinformaticians compared to Nextflow workflows which are written in Groovy.

Strengths:

1. **Flexibility:** Supports a wide range of metabarcoding markers, not limited to microbial and fungal markers.
2. **Scalability:** Designed for use on high-performance computing clusters, facilitating the analysis of large datasets.
3. **Automation:** Utilizes Snakemake for pipeline automation, reducing the need for user intervention.
4. **Taxonomic Assignment:** Uses the RDP Classifier for confident taxonomic assignments and supports multiple markers. Includes the databases for this assignment.
5. **Specialized Processing:** Handles ITS regions by trimming conserved rRNA gene regions and offers pseudogene filtering for protein-coding genes.
6. **Reproducibility:** MetaWorks promotes reproducible research with detailed documentation and versioned software via Snakemake.
7. **User Support:** Provides extensive documentation, a step-by-step tutorial, a FAQ, and quickstart examples for new users.

Weaknesses:

1. **Dependency on Reference Databases:** The accuracy of taxonomic assignments is contingent on the comprehensiveness and quality of reference databases used.
2. **Data Handling Limitations:** Although it's scalable, handling extremely large datasets may require significant computational resources and time.
3. **Specificity to Illumina Reads:** The pipeline is optimized for Illumina sequencing data, which may not be applicable for data from other sequencing platforms without modifications.

Table 1.0: Summary of strengths, weaknesses, and unique aspects of selected metabarcoding pipelines

Pipeline	QIIME 2	Ampliseq	APSCALE	MetaWorks
Main Use	Managing, analyzing, and visualizing microbiome data from DNA sequencing studies.	Processing amplicon sequencing data with support for various gene regions.	Merging paired-end sequences, trimming, filtering, clustering, and denoising for metabarcoding markers.	Metabarcoding data analysis using various markers, optimized for high-performance computing clusters or cloud.

Flexibility	Offers multiple options for analysis. Can be used with python or command line.	Supports a variety of sequencing platforms and amplicon types. Built with Nextflow for portability.	Platform-independent, with a focus on accessibility and ease of use for non-experts. GUI and command-line options available.	Supports a wide range of metabarcoding markers with Snakemake for portability.
Strengths	Comprehensive analysis pipeline, plugin flexibility, community support, interactive visualizations, reproducibility.	Portability across systems, automated quality control and taxonomic classification, reproducibility, and resource-efficient benchmarking capabilities.	Platform independence, user-friendly GUI, scalability, data protection compliance, comprehensive data processing, visualization, and summary statistics in GUI.	Flexibility in markers, scalability, automation, confident taxonomic assignments with curated databases, specialized processing for ITS and protein-coding genes, reproducibility, user support with extensive documentation.
Weaknesses	Steep learning curve, computational resources, frequent updates, variable plugin quality, extensive but overwhelming documentation.	Setup learning curve, computational resources, reduced QIIME 2 flexibility without modifications.	Python dependency, manual demultiplexing, additional steps for Windows users, local data storage space requirements.	Dependence on reference databases, data handling limitations for very large datasets, specificity to Illumina data.
Unique Aspects	The most extensibility, proven research use, strong user and developer community.	Nextflow workflow management system, easy deployment on cloud platforms.	Data protection focus with local run capability, GUI for biologists without bioinformatics background.	Snakemake workflow management system, the familiarity of Python code, optimized for Illumina data, includes specialized processing and filtering capabilities.

2.3 Base calling

Often an overlooked step, base calling is the process of converting raw signal data from sequencers into nucleotide sequences (typically in FASTQ format), and it is a critical and often first step in any bioinformatics workflow. Different sequencing platforms use different methodologies for reading DNA sequences, so the algorithms for base calling are platform-specific. In this section of the guidelines, we review the available base calling software for popular sequencing platforms, such as Oxford Nanopore

Technologies (ONT), Illumina, and Pacific Biosciences (PacBio). We discuss their strengths and weaknesses and provide guidelines for selecting the most appropriate software package for specific use cases. Generally speaking, the software provided by the companies supplying the DNA sequencing instruments is sufficient.

2.3.1 Oxford Nanopore Technologies

ONT sequencers generate long reads in real-time, making them well-suited for de novo genome assembly, metagenomics, and transcriptomics applications (Jain et al., 2016). The two main base calling software options for ONT are Guppy and Bonito.

2.3.1.1 Guppy

Guppy is the default base caller provided by ONT and uses a recurrent neural network (RNN) to convert raw electrical signals into nucleotide sequences (R. R. Wick et al., 2019). Guppy has a high base calling accuracy (>95%) and is optimized for GPU-based computing, which enables rapid processing of large datasets (R. R. Wick et al., 2019).

Strengths:

- High accuracy
- GPU optimization for faster processing
- Supports various sequencing modalities

Weaknesses:

- Closed-source and commercial, limiting customization options
- Lower accuracy for homopolymeric regions (Loman et al., 2021)

2.3.1.2 Bonito

Bonito is an open-source base caller developed by ONT that employs a deep learning approach (Rang et al., 2020). Bonito has been shown to achieve a higher accuracy than Guppy, particularly for homopolymeric regions (Loman et al., 2021).

Strengths:

- Open-source and customizable (Rang et al., 2020)
- Higher accuracy compared to Guppy (Loman et al., 2021)

Weaknesses:

- Slower processing compared to Guppy (Loman et al., 2021)
- Limited support for different sequencing modalities (Rang et al., 2020)

2.3.2 Illumina

Illumina sequencers generate short, accurate reads, making them ideal for resequencing and variant detection applications (Van Dijk et al., 2018). They are the most dominant form of sequencing in the market and many applications use these sequencers (e.g. metabarcoding). The primary base calling software for Illumina was bcl2fastq and has now been superseded by BCL-Convert.

2.3.2.1 bcl2fastq

bcl2fastq is the previous default base caller provided by Illumina, which converts raw base call (BCL) files into FASTQ format. It offers high accuracy, fast processing, and compatibility with Illumina's various sequencing platforms using both 2- and 4-colour chemistries.

2.3.2.2 BCL-Convert

BCL-Convert is the updated default base caller provided by Illumina, intended to replace bcl2fastq. It is available to download for CentOS 7 and Oracle 8, as well as being the default base caller for the Illumina DRAGEN platform.

Strengths:

- High accuracy
- Fast processing
- Compatible with all Illumina platforms

Weaknesses:

- Closed-source and commercial, limiting customization options
- Limited to Illumina sequencers

2.3.3 Pacific Biosciences

PacBio sequencers generate long reads with high consensus accuracy, making them suitable for de novo genome assembly, structural variant detection, and full-length transcript sequencing (Rhoads and Au, 2015). The primary base calling software for PacBio is the SMRT Analysis software suite.

2.3.3.1 SMRT Analysis

SMRT Analysis is the default software suite provided by PacBio for base calling and downstream analyses. It includes tools for converting raw signal data into high-quality consensus sequences using the Hierarchical Genome Assembly Process (HGAP) (Chin et al., 2013).

Strengths:

- High consensus accuracy
- Integrated with downstream analysis tools
- Compatible with various PacBio platforms

Weaknesses:

- Closed-source and commercial, limiting customization options
- Limited to PacBio sequencers

2.3.4 Overall recommendations

The choice of base calling software depends on the specific sequencing platform and research objectives. For ONT data, Guppy offers higher processing speed, while Bonito provides better accuracy, particularly for homopolymeric regions (Loman et al., 2021). For Illumina data, BCL-Convert is the preferred and official choice, replacing bcl2fastq, and offering high accuracy and compatibility with various Illumina

platforms (Illumina, 2023). For PacBio data, the SMRT Analysis software suite is the recommended option, providing high consensus accuracy and integration with downstream analysis tools (PacBio, 2023).

2.4 Quality control and filtering

Metagenomics and metabarcoding workflows require rigorous quality control and filtering to provide actionable results. The accuracy and reliability of the results are highly contingent upon the quality of the input sequence data. Thus, effective quality control (QC) and filtering processes are indispensable for ensuring accurate downstream analyses. This section of the guidelines aim to highlight the importance of QC in metagenomic and metabarcoding studies, review prominent software tools for sequence data filtering, and provide insights into their respective strengths and weaknesses.

2.4.1 Importance of Quality Control in Metagenomic and Metabarcoding Workflows

Reliable sequence quality is paramount for several reasons, including accuracy, computational efficiency, and reproducibility. With respect to accuracy, low-quality sequences can lead to spurious OTU or ASV detection, incorrect taxonomic assignments, and misleading functional annotations. With respect to computational efficiency, filtering out poor-quality reads can considerably reduce the computational requirements of downstream analyses. This is particularly impactful when it comes to OTU/ASV workflows due to error correction models such as DADA2 scaling poorly with the number of sequences presented, significantly increasing the memory required. And finally, correct QC and filtering allows for reproducibility by ensuring consistent measures across different datasets, and thereby enabling more robust comparative studies.

2.4.2 Quality Control Options for Different Sequence Data

As described in the section on base calling, different sequencing platforms, such as Illumina, PacBio, and Oxford Nanopore Technologies (ONT), generate distinct error profiles, necessitating specific QC and filtering approaches to ensure quality results. Here we provide a broad overview and comparison of the available tools for QC and filtering based on the sequencer type.

2.4.2.1 Illumina Data

Illumina platforms, producing relatively short-read sequences, are characterized by systematic errors, like substitution errors, especially towards the ends of the reads.

Software options for quality control of Illumina data:

- **FastQC**: Provides an overview of sequence data quality.
- **Trimmomatic**: Tailored for trimming Illumina sequence data, removing adapters and low-quality bases.
- **BBduk**: Another powerful trimming tool, allowing for more flexibility and customization.
- **Cutadapt** is specifically designed for adapter and quality trimming of short-read sequencing data from platforms like Illumina.
- **Fastp** is another powerful tool for Illumina data QC created by OpenGene and written in C++, known for it's quick processing speed.

Table 1: Comparison of available QC and filtering software for Illumina data

Feature	FastQC (Andrews, 2010)	Trimmomatic (Bolger et al., 2014)	BBduk (Bushnell, 2014a)	Cutadapt (C. Martin, 2011)	FastP (S. Chen et al., 2018)
Adapter removal	No	Yes	Yes	Yes	Yes
Quality filtering	Basic	Advanced	Advanced	Advanced	Advanced
Flexibility	Low	Medium	High	High	High

Strengths and Weaknesses:

- **FastQC** is user-friendly but only provides a diagnostic overview.
- **Trimmomatic** is commonly used for its efficiency, but its less flexible nature can limit certain workflows.
- **BBduk** offers high customizability but has a steeper learning curve.
- **Cutadapt** is highly precise in detecting and removing adapter sequences, which can often contaminate downstream analyses. While it can be used for long-read data like PacBio or Nanopore, it's optimized for short-read data.
- **FastP** provides a comprehensive QC including error rate, GC content, and more. Additionally, it can generate intuitive visualizations for easier interpretation of data quality. It's more tailored towards standard Illumina libraries and might not handle non-standard libraries with the same assurance.

2.4.2.2 PacBio Data

PacBio's Single Molecule, Real-Time (SMRT) sequencing produces long reads but has a higher error rate, primarily insertions and deletions.

Software options for PacBio data:

- **SMRT Analysis**: The native software suite designed for QC and analysis of PacBio data.
- **Canu**: A genome assembler that also offers error correction for PacBio data (Koren et al., 2017)

Table 2: Comparison of available QC and filtering software for PacBio data

Feature	SMRT Analysis	Canu
Error correction	Yes	Yes
Assembly	No[1]	Yes

[1] PacBio's website claims that SMRT Analysis can perform *de novo* assembly but we were unable to find any publications using this feature so we are not able to evaluate its abilities.

Quality visualization	Yes	Limited

Strengths and Weaknesses:

- **SMRT Analysis** provides a comprehensive suite of tools but might be excessive for pure QC purposes.
- **Canu** is versatile but can be resource intensive.

2.4.2.3 Oxford Nanopore Technologies (ONT) Data

ONT sequencing, like PacBio, provides long reads but has unique error profiles, mainly substitutions.

Software options for ONT data:

- **NanoPlot**: Offers quality visualization for ONT data (De Coster & Rademakers, 2023)
- **Porechop**: Used for adapter trimming (R. Wick, 2018)
- **Medaka**: Produced by Oxford Nanopore itself, offers consensus calling and variant calling.

Table 3: Comparison of available QC and Filtering Software for ONT Data

Feature	NanoPlot (De Coster & Rademakers, 2023)	Porechop (R. Wick, 2018)	Medaka (Medaka, 2018)
Adapter removal	No	Yes	No
Quality visualization	Yes	No	No
Consensus/Variant calling	No	No	Yes

Strengths and Weaknesses:

- **NanoPlot** is excellent for data visualization but doesn't provide modification tools.
- **Porechop** is efficient for adapter removal but lacks comprehensive QC features.
- **Medaka** is powerful for improving sequence accuracy but is not a standalone QC tool.

2.4.3 Considerations when Choosing Software Packages

The choice of software should be predicated on the sequencing platform used, specific QC needs, and computational resources. Here are some general considerations:

1. **Error Profile**: Different tools are optimized for distinct error profiles, so understanding the nature of errors in the dataset is crucial.
2. **Usability**: While flexibility and customizability are valuable, they might come at the cost of user-friendliness.

3. **Integration**: Tools that seamlessly integrate with popular downstream analysis pipelines are advantageous.

2.4.4 Overall recommendations

Robust quality control is pivotal in harnessing the full potential of metagenomic and metabarcoding studies. Given the rapidly evolving landscape of sequencing technologies and associated software tools, continuous benchmarking and evaluation of QC tools are imperative. Selecting the appropriate software package necessitates a comprehensive understanding of the sequencing data's unique characteristics and the specific requirements of the research question. Here is a list of general recommendations based on sequencer type.

Table 4: Overall recommendations for each sequencing platform

Sequencing Platform	QC and Filtering Tool	Strengths	Weaknesses
Illumina	Cutadapt (C. Martin, 2011)	Accurate adapter trimming, user-friendly, open-source	Designed primarily for Illumina, might not be ideal for long-reads
Illumina	Fastp (S. Chen et al., 2018)	Comprehensive QC, ultra-fast, provides visualizations	Limited to Illumina data, not ideal for non-standard libraries
Nanopore	Porechop (R. Wick, 2018)	Designed for long-reads, adapter trimming	Less robust for quality filtering
PacBio	SMRT Analysis	Long-read correction, assembling, trimming	Compressive list of features presents overhead when only used for QC

2.5 Denoising

Single nucleotide variants (SNVs) can be caused by a number of factors such as PCR or sequencer artifacts. With traditional alignment-based methods in the context of whole genome sequencing, SNVs are unlikely to drastically alter the accuracy of the alignment. However, with targeted sequencing methods where multiple similar sequences are being compared and assigned taxonomy, these small errors have the potential to lead to incorrect taxonomic assignment. Fortunately, methods to account for this have been developed, most notably clustering similar sequences into “operational taxonomic units” (OTUs) and removing errors through denoising into “amplicon sequence variants” (ASVs). This section of the guidelines will give a broad overview of the topic, as well as offer software and workflow recommendations.

2.5.1 OTUs vs ASVs

OTU clustering is a method used to minimize sequencer errors in targeted sequencing. The OTU approach overcomes PCR and sequencing errors that may be present within metabarcoding datasets by clustering together highly similar sequences (for example, with >98% sequence identity), with the most dominant sequence from each cluster used for taxonomic assignment. OTUs may correspond to ecological species, niche uniqueness, or biological species (i.e., unique reproductive pools) (Cristescu, 2014). There are three types of OTU clustering (Edgar, 2017). The first one, *de novo* clustering, is computationally complex and must be repeated when data are added or removed from the study. The

second one, closed-reference clustering, is more efficient and uses a reference database of target gene sequences from known taxa. However, because it is dependent on reference sequences it can be biased and reads that do not closely match a reference sequence will be discarded. The third method, open-reference clustering, is a combination of *de novo* and closed-reference clustering and avoids the loss of novel sequences.

In contrast to OTUs, ASVs keep each unique sequence separate but filter out potential PCR and sequencing errors based on built-in error models. While overall ecological patterns derived from metabarcoding data tend to be fairly robust to the choice of approach (Glassman & Martiny, 2018), ASVs are more reproducible, and therefore more cross-comparable where linking relies on sequence identity rather than species names (Callahan et al., 2017).

There is considerable debate in the field regarding which method is more appropriate. Some papers have suggested that the field should be moving towards an ASV approach due to its ability to provide a more precise identification of species and a more detailed picture of diversity within a sample (Callahan et al., 2017; Porter & Hajibabaei, 2020). However, other authors have suggested the picture is more nuanced, and each method carries its own trade-offs (Chiarello et al., 2022). While OTUs are relatively computationally fast and easy for both generation and comparison between samples and studies, they can carry a significant risk of reference bias and loss of novel sequences when using closed-reference clustering. Moreover, clustering too coarsely can merge reads from closely related species into single OTUs, leading to ambiguous matches when assigning taxonomy. For this reason, the clustering thresholds for OTU generation need to be empirically established in a marker- and taxon-specific context (Alberdi et al., 2018). ASVs, on the other hand, are computationally slow but will retain all sequences from the sample and have no risk of reference bias as they are generated reference-free.

2.5.2 Software Packages for generating OTUs

Operational Taxonomic Unit (OTU) clustering is a cornerstone of many bioinformatics pipelines, especially in metabarcoding and metagenomics research. The process involves grouping similar sequences together that likely originate from the same species or a set of closely related species. The selection of an appropriate software tool for OTU clustering is essential for ensuring accuracy, precision, and computational efficiency in the bioinformatics workflow.

2.5.2.1 QIIME/QIIME2 (Quantitative Insights Into Microbial Ecology)

QIIME is a complete pipline that arose from the study of microbiome data, but is also applicable to metazoan environmental genomics. It is modular, allowing new capabilities to be introduced into the pipeline through plugins, of which there are already a sizable number that have been contributed by the community (see: https://library.qiime2.org/plugins/).

Strengths:

- Comprehensive tool with end-to-end pipeline from raw sequence data to microbial community analyses.
- Highly customizable and extensive community support.
- Integration with other tools and databases.

- Wraps other clustering algorithms such as Mothur and UCLUST

Weaknesses:

- Learning curve due to its vast set of features.
- Heavy computational resources may be required for large datasets.

Reference: (Bolyen et al., 2019; Caporaso et al., 2010)

2.5.2.2 Mothur

Like QIIME, Mothur is a pipeline that arose from the microbiology community but it is also applicable to metazoan metabarcoding data. Unlike QIIME, it is a single monolithic software program where most of the development has been performed by a single individual and development has slowed since 2022, although at the time of writing the most recent release was from May 2024 (version 1.48.1).

Strengths:

- Open-source and offers a comprehensive suite of tools.
- Most highly-cited software package for 16s rRNA analysis.
- Highly flexible with scriptable commands.

Weaknesses:

- Can be computationally intensive on large datasets.
- The command-line interface might be challenging for beginners.

Reference: (Schloss et al., 2009)

2.5.2.3 UCLUST (part of USEARCH)

UCLUST is part of the USEARCH package, which is commonly used in the academic community because of the free licensing for educational use. However, for commercial use the software is not free.

Strengths:

- Fast and efficient, suitable for large datasets.
- High accuracy in clustering.
- Convenient for high-throughput sequencing data.

Weaknesses:

- Proprietary for the full version.
- Free only for academic use; commercial use requires a paid license.
- Limited features compared to comprehensive tools like QIIME2 and Mothur.

Reference: (Edgar, 2010)

2.5.2.4 VSEARCH

VSEARCH is an open-source alternative to UCLUST/USEARCH; the algorithms have been re-implemented from the descriptions in the original papers. The output is similar to but not exactly the same as the official USEARCH package.

Strengths:

- Open source so the code can be audited and modified if desired.
- Handles noise well and offers dereplication, sorting, and chimera detection.
- Capable of working both 32-bit and 64-bit processors.

Weaknesses:

- Slower than UCLUST/USEARCH on large datasets.

Reference: (Rognes et al., 2016)

2.5.3 Comparative Analysis of Software Packages

2.5.3.1 Performance and Scalability

For large datasets, UCLUST stands out for its efficiency, followed by VSEARCH. QIIME2 and Mothur can handle comprehensive analyses, but their performance might be a limiting factor for exceptionally large datasets.

2.5.3.2 Accuracy

All tools have shown comparable accuracy, but the precision can vary based on the type and quality of input data. Mothur and QIIME2 are often highly regarded due to their comprehensive nature and detailed protocols that guide users through the quality control steps.

2.5.3.3 Flexibility and Integration

QIIME2's integration with various databases and tools makes it a one-stop solution for many researchers. Mothur's scripting capabilities provide a high level of flexibility for custom analyses.

2.5.3.4 User Interface and Learning Curve

QIIME2, being comprehensive, has a steeper learning curve. Mothur's command-line interface is robust but may be intimidating for novices. UCLUST and VSEARCH are more straightforward in their function but might require integration with other tools for a complete pipeline.

2.5.4 Best Practices for OTU Clustering

1. **Quality Control**: Always filter and trim raw sequences to remove low-quality bases, adaptors, and potential contaminants.
2. **Dereplication**: Identifying sequences that are *identical* to each other and removing duplicates to reduce the size of the dataset.
3. **Choose Appropriate Similarity Threshold**: Ideally this should be equal to or less than the intraspecific variation for the particular gene segment that was sequenced, but greater than the

inter-specific distance. A 97% similarity threshold is often used for bacteria and archaea, but this can vary depending on the research question.

4. **Check for Chimeras**: Chimeric sequences (see the Glossary) can distort downstream analyses. Tools like UCHIME (part of USEARCH and VSEARCH) or those built into Mothur and QIIME2 can be used for chimera checking.
5. **Consider the Nature of Data**: For instance, if working with ITS sequences, consider software specifically designed for such data due to its variability.
6. **Computational Resources**: Ensure you have access to the necessary computational power, especially for large datasets. Some tools can be parallelized or run on high-performance computing clusters.

Table 5: summary of strengths and weaknesses for OUT clustering packages

Software	Strengths	Weaknesses
QIIME2 (Bolyen et al., 2019)	- Comprehensive tool. - Highly customizable. - Integration with other tools, algorithms, and databases.	- Learning curve. - Needs heavy computational resources for large datasets.
Mothur (Schloss et al., 2009)	- Open-source with comprehensive tools. - Responsive community. - Flexible scripting.	- Computationally intensive on large datasets. - Command-line interface may be challenging for beginners.
UCLUST (Edgar, 2010)	- Fast and efficient. - High accuracy in clustering. - Good for high-throughput sequencing data.	- Proprietary and license fees are required for the full version - Limited to 32bit for the free version.
VSEARCH (Rognes et al., 2016)	- Open-source alternative to UCLUST. - Handles noise and offers chimera detection.	- Slower than UCLUST on large datasets. - Not as well tested as UCLUST from USEARCH

2.5.5 Software Packages for generating ASVs

Rather than grouping sequences into operational taxonomic units based on arbitrary thresholds, ASVs aim to resolve sequences at single-nucleotide resolution. This offers more accurate and reproducible results. In this section, we will evaluate various software options available for generating ASVs, emphasizing their strengths, weaknesses, and applicability to different types of data.

2.5.5.1 DADA2

Description:

DADA2 is a model-based approach for correcting sequencing errors in Illumina data, providing single-nucleotide resolution.

Strengths:

- **Accuracy**: DADA2 tends to have lower error rates in benchmarking studies compared to other methods.
- **Accuracy for rare species**: It can differentiate between very closely related taxa.
- **Integrated Workflow**: DADA2's R package includes functions for quality filtering, dereplication, and taxonomic assignment.

Weaknesses:

- **Computationally Intensive**: Requires significant computational resources for larger datasets.
- **Illumina-specific**: Optimized for Illumina data and might not perform as well with other platforms.

Reference: (Callahan et al., 2016; Nearing et al., 2018)

Applicability:

Optimal for Illumina amplicon datasets, especially when high-resolution taxonomic differentiation is required.

2.5.5.2 Deblur

Description:

Deblur employs a de-novo approach to obtain ASVs resolution by removing sequencing errors.

Strengths:

- **Speed**: Faster than DADA2, particularly with large datasets.
- **Resolution**: Offers sub-OTU resolution, useful for identifying closely related organisms.
- **Noise Reduction**: Uses a known error model to remove noise from the dataset.

Weaknesses:

- **Strict Quality Filtering**: Rather than correcting identified errors, it eliminates reads that are determined to contain errors. This can result in the loss of a significant portion of the data.
- **Dependency**: Dependent on QIIME2 for full pipeline functionality.

Reference: (Nearing et al., 2018)

Applicability:

Best suited for Illumina datasets where speed is a priority, and researchers are working within the QIIME2 environment.

2.5.5.3 USEARCH – UNOISE3

Description:

UNOISE3 is a part of the USEARCH suite and provides error correction to generate ASVs.

Strengths:

- **Speed**: It's a faster method, particularly noticeable in large datasets.
- **Less Stringent Quality Filtering**: Potentially retains more data than Deblur.
- **Broad Applicability**: Can be used with multiple sequencing platforms.

Weaknesses:

- **Less Resolution**: Although it captures ASVs, it may not provide as high resolution as DADA2.
- **License Requirement**: USEARCH is not open-source, and a license is required for commercial use for the 64-bit version.

Reference: (Antich et al., 2021; Nearing et al., 2018)

Applicability:

For researchers with datasets from various sequencing platforms and those already familiar with the USEARCH ecosystem.

2.5.5.4 VSEARCH – UNOISE3

Description:

Open-source implementation of UNOISE3 included as part of the VSEARCH package.

Strengths:

- **Open-source**: Freely available without licensing fees.
- **Chimera Removal**: Includes an open source implementation of the uchime3_denovo algorithm as well.
- **Broad Applicability**: Can be used with multiple sequencing platforms.

Weaknesses:

- **Validation**: Most papers and researchers are using UNOISE3 from USEARCH, this alternative implementation may not be as well tested.

Reference: (Rognes et al., 2016)

Applicability:

For researchers with large datasets who are unable to access the 64-bit version of USEARCH.

2.5.6 Best Practices for ASV denoising

2.5.6.1 Quality Control and Filtering

Before denoising, raw sequences should undergo quality control:

- Use tools like FastQC or MultiQC to assess the quality of your raw data.
- Trim or remove low-quality bases from the ends of sequences. Programs like Cutadapt or Trimmomatic can be employed.
- Filter out sequences below a certain quality score threshold.
- Remove any non-biological sequences, such as adapters or primers.

2.5.6.2 Optimize Parameter Choices

Most ASV denoising tools offer various parameters that influence the denoising process. Depending on the dataset's characteristics, researchers might need to:

- Adjust error rate parameters. This can prevent over- or under-clustering, controlling the number of false positives and negatives.
- Customize length trimming parameters, especially if amplicon lengths are variable.

2.5.6.3 Track and Visualize Denoising Metrics

- Monitor the number of sequences retained or discarded during each step of the denoising process. Sudden and significant drops can indicate issues that require troubleshooting. For example, if the target amplicon is ~250 bases in length and the read length for forward and reverse read sequencing is 150 bases, one would expect that the vast majority of reads will merge successfully. If a large proportion of a sample's reads are lost at the merging step, it could be an indication that a high amount of off-target DNA was sequenced.
- Utilize visualization tools like those provided within QIIME2 or R packages to inspect denoising results.

2.5.6.4 Consider Biological Context

While denoising software relies on mathematical models to reduce errors, always keep the biological context in mind:

- Does the data contain closely related taxa? If so, high-resolution methods like DADA2 may be particularly valuable because it can perform precise error correction at the individual nucleotide level. Conversely, OTU clustering is typically performed at a sequence similarity of 98-99% which could create ambiguous matches within some taxonomic groups (e.g., some members of the Salmonidae fish family are very difficult to distinguish using common DNA barcoding markers because the sequence similarity is so high.
- Be wary of discarding rare ASVs outright, as they could represent legitimate low-abundance organisms. Ironically, these may be of most interest (e.g., rare or endangered species).

2.5.6.5 Regularly Update Software

- Denoising algorithms and software tools undergo regular updates which can introduce new features, optimizations, or bug fixes.
- Ensure you are using the latest stable version of your chosen software.
- Review changelogs or release notes for significant updates or parameter changes that might affect your analyses, or may affect comparisons with results generated using previous software versions.

2.5.6.6 Denoising with Different Sequencers

While many denoising tools are optimized for Illumina data, researchers using other platforms should:

- Seek tools or parameters specifically designed for those platforms, such as Medaka for Oxford Nanopore Technologies (Medaka, 2018).
- Consider cross-referencing with platform-specific forums or communities for insights into platform-specific quirks or challenges.

2.5.6.7 Documentation and Reproducibility

- Thoroughly document every step of your denoising process, including software versions and parameter choices.
- Whenever possible, use workflow management tools like Snakemake (Köster & Rahmann, 2012)or Nextflow (Di Tommaso et al., 2017) to ensure reproducibility.

2.5.6.8 Benchmarking and Validation

- If possible, include a mock community with known composition in your sequencing run. This allows you to validate the denoising process and assess the accuracy of your ASVs.
- Compare denoising results from multiple software options to gauge consistency and potential biases.

Table 6: Summary of strengths and weaknesses for ASV denoising packages

Software	Strengths	Weaknesses
DADA2 (Callahan et al., 2016)	- Comprehensive R software package. - Highly customizable. - Accurate for rare species detections	- Need R programming knowledge if not using a wrapper - Needs heavy computational resources for large datasets - Slower than other options - Additional workarounds may be needed when working with quality binned data

Deblur (Nearing et al., 2018)	- Open-source - Fast and efficient - Integrated with QIIME2	- May discard more data than other options - Command-line interface outside of QIIME2 has not been updated for some time
UNOISE3 (Edgar, 2010)	- Fast and efficient. - Good for high-throughput sequencing data - Handles denoising and offers chimera detection via uchime3_denovo. - Works well with various sequencer types, such as the NovaSeq	- Proprietary and license fees are required for the full version - Limited to 32-bit for the free version.
VSEARCH (Rognes et al., 2016)	- Offers open-source alternative to UNOISE3 algorithm. - Includes open source alterative implantation of the chimera removal algorithm uchime3_denovo	- Not as well tested as UNOISE3 from USEARCH

2.5.7 Guides for Selected Denoising Software

2.5.7.1 Guide for UNOISE3

UNOISE3 is the latest denoising algorithm by Robert Edgar included in the proprietary software package USEARCH (Edgar, 2010). There is also an open source implementation of the algorithm in the program VSEARCH (Rognes et al., 2016).

The general UNOISE workflow involves truncating reads to the same length (although this step is not strictly necessary when reads are paired-end and merged), merging, dereplicating, running the UNOISE algorithm, and removing chimeras. The ASVs produced by UNOISE are highly dependent on the selection of an alpha parameter. The alpha parameter is a value that controls the shape of the error distribution in the UNOISE algorithm. Specifically, it determines the degree of randomness or variability in the noise introduced by the algorithm. A higher value of alpha leads to a more variable error distribution, while a lower value of alpha leads to a less variable error distribution. In practical terms, a higher value of alpha means that the UNOISE algorithm is more likely to include random sequencing errors as part of the output, which can result in more diverse sequences being detected. However, this can also increase the likelihood of false positives and reduce the accuracy of the resulting data. On the other hand, a lower value of alpha results in a more conservative error model, which can be more accurate but may miss some rare sequences. In general, the choice of the alpha parameter depends on the specific application and the desired trade-off between accuracy and sensitivity. A common approach is to try a range of alpha values and evaluate the results on a validation set to determine the optimal value for the given dataset and research question. A literature review may be conducted as well to determine the recommended alpha parameter for a given dataset. For COI data for example, other papers have observed an alpha parameter of 5 works reasonably well (Antich et al., 2021).

2.5.7.2 Guide for DADA2

DADA2 is a denoising algorithm (Callahan et al., 2016). Comparisons with UNOISE3 and Deblur have shown DADA2 to have the most sensitivity in discerning rare low abundance ASVs. DADA2 has also been shown to be the most resource intensive compared to UNOISE3 and Deblur, with a higher run time and RAM requirements. The general workflow of DADA2 is similar to UNOISE3, truncating reads to the same length, merging, dereplicating, running the DADA2 interface and removing chimeras. As DADA2 is installed as an R package, it contains R functions for these tasks that can be glued together as part of an R script. While in most cases, the run time with DADA2 will be acceptable, with large datasets, it is recommended to delegate tasks such as truncation and merging to more performance optimized applications such as FastP, and then importing the data to run the DADA2 interface. Furthermore, should one wish to run DADA2 pipeline without writing R code, easy to use wrappers such as Dadaist2 can be used to conveniently run DADA2, as well as providing additional features such as the ability to generate QC reports in a format compatible with MultiQC (Ansorge et al., 2021). Special attention should be given to running DADA2 with sequencer types such as the NovaSeq. As the NovaSeq bins quality scores, they do not represent a continuous distribution and this can affect the error calculations.

2.5.8 Novel Denoisers

Some additional denoisers have been created since the introduction of UNOISE3 and DADA2. Some offer unique ideas or user friendly features such as NG-Tax 2.0, which easily plugs-in to the Galaxy ecosystem (Poncheewin et al., 2020). While ideas presented in these novel denoisers can be compelling, newer denoisers lack the battle-tested validation of established denoisers such as DADA2 and UNOISE3 for a wide variety of markers in metabarcoding studies. In addition, future developments for these software packages can be uncertain; software arising from academic labs often becomes abandoned after students graduate or funding dries up. Therefore, it is recommended to use established denoisers for most use cases and only deviate from the established packages if a novel denoiser offers a benefit to the research question that is not offered by existing packages such as DADA2 or UNOISE3.

2.6 Chimera removal

Chimeras occur during the PCR step when two strands of DNA that are not the correct complements of each other become joined together, creating a new amplicon that was not present in the original sample. This capability is built into tools that perform denoising/OTU clustering so the choice of tool will be guided by the choice of tool used in the previous step. Briefly, DADA2 implements the "removeBimeraDenovo" function (Callahan et al., 2016), USEARCH has a "uchime3_denovo" command (Edgar, 2010), and this same command is also implemented in VSEARCH (Rognes et al., 2016).

2.7 Taxonomic assignment

2.7.1 Introduction

If there were no PCR or sequencing errors in metabarcoding sequence reads, and if reference databases were complete and contained exemplars for all intraspecific variation, and these reference sequences were distinct from those from congeneric species, then taxonomic assignment would be an easy task: it's a simple matching exercise. Unfortunately, reality is not so simple, so it is necessary to use algorithms that find the most probable taxonomic assignment for any given query sequence.

In cases where sequencing/PCR errors are minimal and the target species is represented in the reference database and can be unambiguously distinguished from closely-related species, the choice of algorithm is somewhat arbitrary—all will come to the correct conclusion. But this is not the case for the majority of reads from a typical metabarcoding study, and algorithm choice can have an impact on the taxonomic assignments made to the data.

Choice of taxonomic assignment method and taxon acceptance thresholds (i.e., the number or proportion of sequence reads required for an OTU/ASV to be retained in the final dataset) can alter the interpretation of species detections. Optimal parameter choices will depend on the characteristics of the marker used, the completeness of the reference database, and the purpose for which the data is to be used. For instance, if the aim is to assess overall ecological patterns, then more aggressive filtering may be chosen to reduce noise while there is a relatively low cost for inaccurate taxonomic identification. However, if the aim is to detect invasive or endangered species, even very weak detections may be considered, and each species needs to be identified with a high degree of accuracy.

In general, there is a dichotomy in the choice of taxonomic assignment algorithm (Table 7): there are methods that are highly accurate but have a slow execution speed (phylogenetic approaches), those that are fast but tend to have lower accuracy (k-mer based approaches), and those that fall in between in terms of both accuracy and execution speed (sequence similarity-based approaches).

In the following sections we will discuss each of these methods in greater depth and make recommendations about which methods are most appropriate in which situations.

Table 7: A dichotomy exists in taxonomic assignment algorithms. Some algorithms are very fast, but at the expense of accuracy. While other methods are very accurate but can be infeasibly slow to run.

Category	Tool(s)	Speed	Accuracy
K-mers	Kraken2 (D. E. Wood et al., 2019), RDP classifier (Maidak et al., 1996), QIIME2 feature classifier (Bokulich et al., 2018)	Very high	Very low
Similarity search	MegaBLAST (Z. Zhang et al., 2000)	High	Low
	Discontiguous MegaBLAST (Altschul et al., 1997)	Medium	Medium
	BLASTN (Altschul, 2014)	Low	High
Phylogenetic	EPA-NG (Barbera et al., 2019), ProTax (Somervuo et al., 2016)	Very low	Very high

2.7.2 K-mer (machine learning) based approaches

With these techniques, classification is based on the number of k-mers (i.e., a DNA "word" of length *k*) a query sequence has in common with a reference sequence. Then simple Bayesian statistics can be used to evaluate the posterior probability of the number of matching k-mers between a query and a target. A convenience of these models is that they can be trained to make taxonomic assignments at multiple levels. For example, if a species-level match is not possible with the reference database then the model can assign genus, family, or higher orders of taxonomy based on its probabilistic model.

Popular tools that implement this technique include the QIIME2 feature classifier (Bokulich et al., 2018), the RDP classifier (Maidak et al., 1996), and kraken2 (D. E. Wood et al., 2019).

Another advantage of these techniques is that they are very fast when assigning taxonomy; kraken2, for example, can assign taxonomy to thousands of reads per second on standard computer hardware. However, it is very slow and computationally intensive to build the reference databases. Again, in the case of kraken2, a large memory machine (i.e., more than 256GB of RAM) is required to build a reference database from GenBank and the process takes several days. In another example, QIIME2's classifier required more than 500GB of RAM and two days to train a model to identify a relatively small reference database of fish species (Hleap et al., 2021). Pre-computed reference libraries are available which removes this burden. For example, the Midori2 database can be downloaded in formats suitable for use with the QIIME2 and RDP classifiers (Leray et al., 2022). This, combined with the speed of execution of taxonomy assignment makes this method very popular.

Like all machine learning algorithms, a significant drawback to these techniques is that they are highly influenced by their training set. For example, if a model is trained solely on fish COI sequences, it may think that a particular k-mer is unique to a species of fish with 100% confidence. But if it had been trained on a broader dataset, perhaps that k-mer also appears elsewhere in the evolutionary tree—a bacterium or another animal, for example. This is why, ironically, these approaches become less accurate when a larger reference database is used (Nasko et al., 2018).

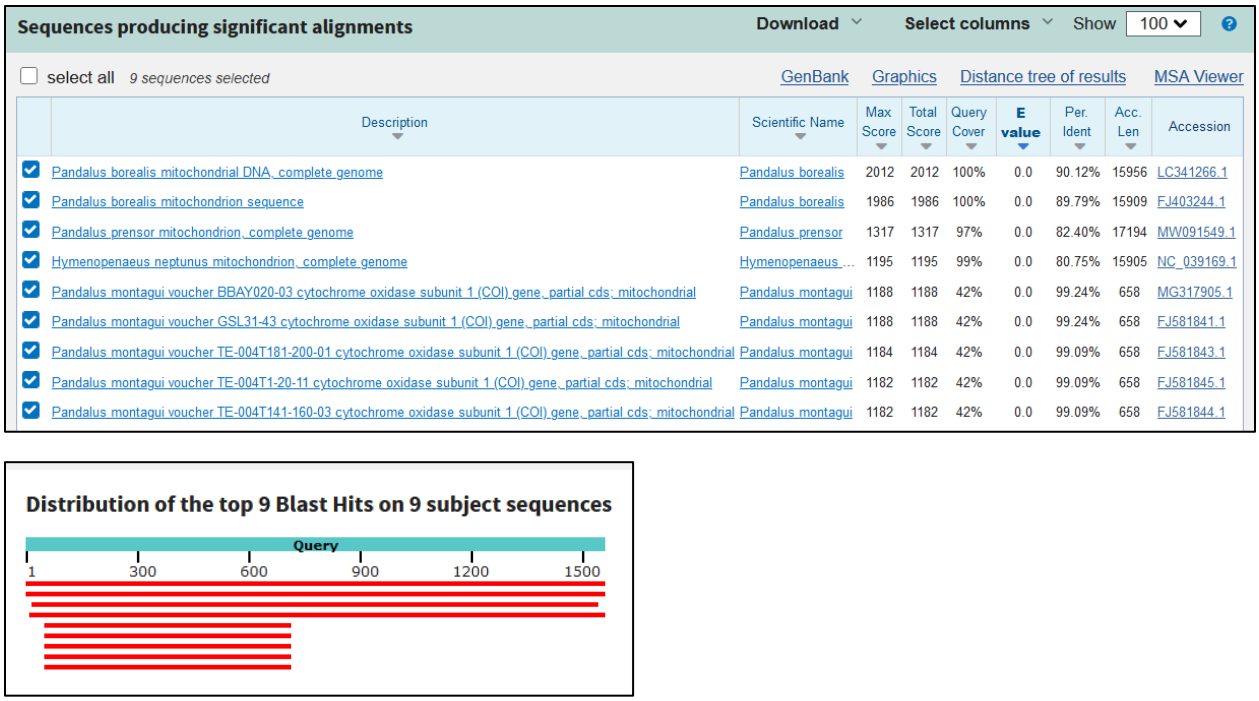
Because of these drawbacks, we recommend using these classifiers only in cases where speed is paramount or as a form of independent verification of a taxonomic assignment performed using a different algorithm.

2.7.3 Sequence similarity

Probably the most common set of taxonomic assignment techniques are based on sequence similarity searches. Here, alignments are performed between the query sequence and all the members of a reference database to find the best match. Because comprehensive reference databases (like GenBank) can be very large, and the number of sequences arising from a metabarcoding study can also be very large, this means many billions of pairwise alignments may need to be performed if a naïve approach were taken. To speed up this process, software tools have implemented heuristics and do not perform complete alignments. The most famous of these tools is BLAST (Altschul, 2014), which pre-processes the query sequence to built a lookup table that enables it to quickly assess whether there exist short, high-scoring ungapped alignments between the query sequence and a potential target. In this way, a large proportion of the reference database can be quickly eliminated from further consideration. In its next step, BLAST attempts to extend these ungapped "seeds" into a fully gapped alignment. By default, nucleotide BLAST (BLASTN) uses 11-mer seeds. An updated algorithm, MegaBLAST (Z. Zhang et al., 2000), sets this seed length at 28 bases by default, leading to a 10x speed-up over BLASTN but at the sacrifice of sensitivity: MegaBLAST is very accurate at finding closely-related targets in the reference database but it is far less accurate at finding more distantly-related sequences. What this means in practice is that species-level matches are reliable, but if the reference database is incomplete and you are trying to make a family- or order-level match to your query sequence then BLASTN is the better choice. Yet another algorithm in the BLAST family is "discontiguous megaBLAST" (Altschul et al., 1997). Here, the "seed" has gaps in it so not every position needs to match exactly. This is particularly useful in coding sequences where the third codon position is highly variable because it can mutate without changing the encoded protein. Discontiguous megaBLAST is slower than megaBLAST but faster than

BLASTN, and it has greater sensitivity than megaBLAST but less so than BLASTN, so it is somewhat of a compromise between these extremes.

One potential problem with alignment-based approaches to assigning taxonomy is illustrated in Figure 2. Here, a full-length COI sequence was queried against GenBank, which does not (yet) contain this sequence. In reporting results, BLAST has prioritized the length of the query sequence that matches a target above the sequence similarity of the two sequences (note that the top hit matches the full length of the query sequence but at only 90% identity). The correct results appear lower, matching only 42% of the length of the query sequence but at nearly 100% identity because these records are truncated COI sequences. There are two main approaches to solving this problem: (1) the BLAST results can be sorted by descending sequence identity rather than the default sort order of descending BLAST score; or (2) the query sequence can be modified so its length matches the most common barcoding amplicons represented in the reference database (i.e., in the case of COI the sequence could be truncated to the first 650 nucleotides).



Sequences producing significant alignments Download Select columns Show 100

select all *9 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Pandalus borealis mitochondrial DNA, complete genome	Pandalus borealis	2012	2012	100%	0.0	90.12%	15956	LC341266.1
Pandalus borealis mitochondrion sequence	Pandalus borealis	1986	1986	100%	0.0	89.79%	15909	FJ403244.1
Pandalus prensor mitochondrion, complete genome	Pandalus prensor	1317	1317	97%	0.0	82.40%	17194	MW091549.1
Hymenopenaeus neptunus mitochondrion, complete genome	Hymenopenaeus ...	1195	1195	99%	0.0	80.75%	15905	NC_039169.1
Pandalus montagui voucher BBAY020-03 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	Pandalus montagui	1188	1188	42%	0.0	99.24%	658	MG317905.1
Pandalus montagui voucher GSL31-43 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	Pandalus montagui	1188	1188	42%	0.0	99.24%	658	FJ581841.1
Pandalus montagui voucher TE-004T181-200-01 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	Pandalus montagui	1184	1184	42%	0.0	99.09%	658	FJ581843.1
Pandalus montagui voucher TE-004T1-20-11 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	Pandalus montagui	1182	1182	42%	0.0	99.09%	658	FJ581845.1
Pandalus montagui voucher TE-004T141-160-03 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial	Pandalus montagui	1182	1182	42%	0.0	99.09%	658	FJ581844.1

Figure 2: Alignment-based approaches can produce erroneous results if used naively. Here, the query sequence is a full-length COI gene for P. montagui. But the BLAST algorithm gives a higher score to full-length matches of closely related species (top four hits) than to shorter COI fragments from the correct species (the bottom five hits).

Another consideration with similarity searches is that one should generally not simply choose the "best hit" from the results. There will almost always be a "best hit", even if the similarity is not particularly high. If the query sequence's best match to a target has only 80% similarity, for example, the evolutionary distance is clearly too great to be considered a species-level match. To solve this problem, a "lowest common ancestor" approach can be taken: a set of best matches is gathered and then consensus among these top hits is used to assign taxonomy at different levels. Unfortunately, the threshold for what set of matches is "good enough" to be part of the consensus set is not easily determined because it's dependent on the intra-specific variation within a particular marker for a

particular taxon. The level of consensus required is also somewhat arbitrary. It is good practice to repeat the analysis for a subset of the data over a range of parameters to estimate the stability of the species assignments being made. QIIME2 (Bolyen et al., 2019) has built-in plugins that can generate consensus taxonomy from either BLAST or vsearch (Rognes et al., 2016) results. A potential pitfall with this technique is database bias: records from population genetics studies may lead to dozens or hundreds of sequences from the same species being deposited into public databases, which will bias the pool of results from which the consensus is derived. One way to fix this is to have only one exemplar sequence from each species in the reference database, similar to NCBI's RefSeq database (Pruitt et al., 2007).

2.7.4 Phylogenetic approaches

Sequence similarity approaches simply look at the overall difference between the query sequence and a target, but they do not consider *where* those differences occur within the context of sequence variation across all similar sequences. This is where phylogenetic approaches are much stronger. They take an alignment of input sequences and create an evolutionary model that best explains the relationship between these sequences. It can then place the query sequence in the position within the evolutionary tree that has the greatest likelihood, according to this model (Figure 3).

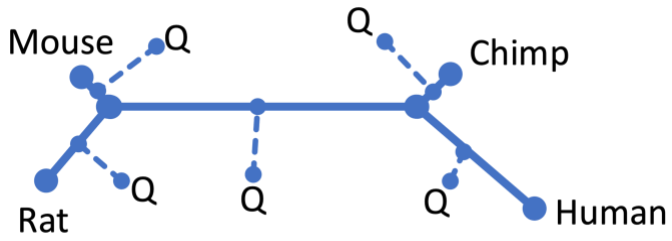


Figure 3: Reference sequences are used to build a phylogenetic tree using maximum likelihood methods (here, composed of mouse, rat, chimp, and human sequences). Then all possible placements of the query sequence (Q) on this tree can then be evaluated within the context of the evolutionary model, and an appropriate taxonomic assignment can be made.

There is little doubt that phylogenetic approaches to taxonomy assignment are the most accurate. An additional benefit is that it is possible to use phylogenetic tools like bootstrapping to estimate the confidence of the assignment of a query sequence to a particular part of the taxonomic tree. Unfortunately, these techniques are also the most computationally intensive which is largely why they are not used more frequently. The authors of EPA-NG, the most advanced software for this technique, used a 2048-core supercomputer to assign taxonomy to 1 billion reads and this took 7 hours—and this is with the benefit of having a reference tree to map against rather than having to generate a reference tree from scratch (Barbera et al., 2019). For context, an S4 flow cell for the Illumina NovaSeq 6000 can produce 8-10 billion reads per run. On a dataset this size, taxonomic assignment can be performed on the order of minutes to hours using k-mer methods, hours to days using sequence similarity searches, and years (!) using phylogenetic methods.

Despite these shortcomings, phylogenetic techniques can be a useful part of a tiered approach to taxonomic assignment. For example, one could use BLAST to assign taxonomy to the majority of reads, then phylogenetic approaches could be used to attempt to assign taxonomy to the subset of reads for which BLAST returned ambiguous results to see if these ambiguities can be resolved with a more sophisticated model.

2.7.5 Recommendations when accuracy is paramount

If a well-curated and complete reference database is used, any of the above methods will arrive at the correct result. However, this ideal situation is rarely possible, so if accuracy is important then the following approach can help ensure the minimization of misclassifications:

- Use the most accurate algorithm possible that can run within an acceptable timeframe (see Table 7).
- Experiments should be run with multiple markers, perhaps from different genetic compartments (e.g., COI from the mitochondrial genome and 18S rRNA from the nuclear genome), and only accept taxonomic assignments where there is agreement between two or more markers.
- If a species-level match is not possible for a particular read, re-analyse the data using different algorithms as independent validation for deeper taxonomic assignments (i.e., family or order-level taxonomy)

3 Reference databases

3.1 Overview

In a metabarcoding study it is usually desirable to match DNA sequences to the individual species they originated from (although so-called "taxonomy free" methods of ecosystem assessment are also valuable (Section 4.2.2.2.2). Thanks to large global initiatives like the International Barcode of Life project (Adamowicz, 2015), standard DNA barcode markers exist for tens of thousands of species and the reference database is growing at a rate of approximately 50% annually (Porter & Hajibabaei, 2018).

Choosing an appropriate reference library can have a dramatic impact on the proportion of sequence reads for which taxonomy can be assigned, and the quality of those assignments (Figure 4). In this section, the pros and cons of different reference databases will be discussed.

Improvements in taxonomic resolution using eDNA requires robust reference DNA databases (Baker et al., 2018; Everett et al., 2018; Hebert et al., 2003a). Reference DNA databases, such as BOLD and GenBank, match curated and verified species to their genotypic sequences (Cowart et al., 2015; Thomsen et al. 2016) and the databases grow as more studies are conducted to contribute verified species. Although databases entries are currently lacking for many species, reference libraries in general are growing (Ratnasingham et al., 2007; Vitecek et al., 2017) and will improve over time. However, many taxonomic groups have little to no representation within reference libraries, and as such taxonomic assignment from a metabarcoding dataset is currently limited.

Limitations such as insufficient DNA reference libraries could potentially be avoided by skipping the species identification step and assigning ecological values to molecular operational taxonomic units (MOTUs), correlating between species/OTUs occurrence and environmental factors (Apothéloz-Perret-Gentil et al., 2017). This type of taxonomy-free molecular index was tested using a diatom index and showed higher correlation between morphological and molecular indices without taxonomic assignment (Apothéloz-Perret-Gentil et al., 2017). An additional benefit of this approach is that it used 95% of the OTUs in the eDNA sample, as opposed to 35% with the taxonomic assignment approach.

Figure 4: Curated reference databases (top) provide the greater accuracy than uncurated databases (bottom), but at the expense of the number of taxonomic assignments made within the dataset.

3.2 Choosing an appropriate reference database

The most comprehensive public reference dataset available are the records within the GenBank/EMBL/DDBJ consortium. However, the data are user-submitted with minimal quality checking, so errors are present. Most of these arise from the submitter having an incorrect species identification. Moreover, the original specimen from which the DNA sequence was derived may not be available for examination, so it may be impossible to independently verify the identification.

In general, there is a trade-off between the size of the reference database and the quality of information within it (Figure 4). Curated databases will return a smaller number of hits but with greater certainty that the correct species assignment has been made. Uncurated databases, on the other hand, will provide species-level matches for a greater proportion of sequences, but with less trustworthy results. Ultimately this will affect the balance between false positives and false negatives in the analysis. Curated databases will have more false negatives but fewer false positives, while more comprehensive uncurated databases will have fewer false negatives but more false positives.

An additional consideration is that, depending on the methodology used for taxonomic assignment (see Section 2.7), larger reference databases can greatly increase the length of computational time required.

An important risk to recognize when employing customized databases is that you will only match what you *expect* to see. For example, if you limit your reference database to a certain reference marker then you could miss important laboratory problems such as primers that are amplifying off-target genes. If you use a reference database that is limited by geography, then you will fail to detect species that are unexpected, such at those introduced through laboratory or field contamination, species that have been recently introduced (through shipping activities, for example), or migration of species into new territories because of changing global climate trends. This can obviously lead to false negatives, but false positives are also possible if two closely related species are present in an environment but only one is represented in the reference database.

Here, we recommend that for general biodiversity studies where gathering the broadest amount of species data possible is the goal, it is best to use the largest reference databases available (e.g., GenBank). In contrast, in applications where accuracy is paramount (e.g., tracking invasive species, rare/endangered species) it is important to use a highly curated database that can stand up to independent verification and the scrutiny of a legal challenge.

3.3 Uncurated databases

The largest publicly available collection of reference sequences for metabarcoding is contained within the GenBank/EMBL/DDBJ consortium (Benson et al., 2018). It is important to recognize that even though this is the largest database available, it still represents only a small fraction of known species (see Section 3.6).

GenBank records are user-submitted with very little quality checking and therefore errors exist. Unfortunately, based on the personal experience of CEGA staff and their connections within the bioinformatics community, it is often difficult to get GenBank staff to correct errors when they are detected. GenBank is understaffed so queries will frequently go unanswered. Moreover, their process involves contacting the original submitter of the record to ask them to correct or withdraw the record, and those individuals may themselves be unresponsive. Tools like BLAST (Altschul, 2014) allow you to mask certain records from being returned in search results, so it is a common practice to keep a local list of GenBank accession numbers that are deemed to contain errors and use this list to filter the results returned from BLAST searches.

Given the low amount of oversight, it is perhaps surprising that the error rate of GenBank has been determined to be fairly low, probably <1% at the genus level (Leray et al., 2019). However, while this may be deemed to be an acceptable error rate for a baseline biodiversity survey, for certain applications where accuracy is paramount (e.g., the detection of endangered species), this error rate may not be acceptable.

There are some ways to ameliorate errors within GenBank and these are described at greater length within the section of this document that focuses on taxonomic assignments (Section 2.7).

3.4 Curated databases

There are many curated DNA barcode reference databases—indeed, too many to provide an exhaustive list here, although we present an example subset in Table 8. Databases are usually focused on specific markers, groups of organisms, geographic region, or a combination of these factors. Unfortunately, many of these reference databases were developed by academic groups for a one-off project and may not be actively maintained. Due to the rapidly growing set of public DNA barcode records, these curated databases can become quickly out of date. It is highly recommended that the date of the last update of a curated reference library is taken into account before deciding to use it.

Table 8: Examples of curated DNA barcode reference databases

Name	Genetic marker	Taxonomic group	Geographic region	Reference
RefSeq	All	All	All	(Pruitt et al., 2007)
Midori2	All	All	All	(Leray et al., 2022)
COInr	COI	All	All	(Meglécz, 2022)

Meta-fish-lib	Multiple	Fish	UK	(Collins et al., 2021)
Silva	16S/18S rRNA	All	All	(Quast et al., 2013)
RDP	16S/18S rRNA	All	All	(Cole et al., 2009)
BOLD[2]	Primarily COI	All	All	(Ratnasingham & Hebert, 2007)
Coins	COI	Insects	All	(Magoga et al., 2022)
FishCard	12S	Fish	California	(Gold et al., 2020)
PLANiTS	ITS	Plants	All	(Banchi, Ametrano, et al., 2020)
GTDB	N/A (whole genome)	Prokaryotes	All	(Parks et al., 2022)

3.5 Creating a custom database

As noted above, many custom databases were assembled for a specific project and are not continually updated. Therefore, if a curated database is desired then it is often preferable to build one *de novo*. There are many tools to assist with this process, many of which also contain functions to assist in identifying and removing bad records.

There are several caveats to be aware of when downloading a subset of data from public databases such as GenBank. The first is that gene annotations are not standardized, and therefore a naïve search query will miss many records. For example, the most common DNA barcode marker for animals is the 5' end of the mitochondrial cytochrome oxidase subunit I gene. However, this is variously annotated as "COI", "CO1", and "COX1" within GenBank, as well as various forms of its full-length name. For this reason, many reference databases like Midori2 are build based on sequence similarity searches rather than searching record metadata (Leray et al., 2022). Another important caveat is that the GenBank taxonomy database differs from other authoritative taxonomic sources such as the Global Biodiversity Information Facility (Telenius, 2011). In a recent study, 26,900 Arthropod entries had a discrepancy in family-level assignment between GBIF and GenBank (Veldsman et al., 2022). Therefore, taxonomy-based searches in GenBank may not always produce the expected results. Some reference database curation tools are aware of these taxonomic synonyms (see below).

If the QIIME2 pipeline is being used, a detailed step-by-step workflow for creating a QIIME2-formatted reference database is available (Dubois et al., 2022). There are several software packages that can facilitate the creation of a custom database for use with a variety of taxonomic annotation tools (Table 9). Because this is an area of rapid development, this list is likely to change rapidly in the coming years but it serves as a guide for what features these packages will typically provide. Each of these software

[2] BOLD also contains unverified records but they are annotated as such

packages has its pros and cons, but we will highlight some features of a few of them. BCdatabaser is perhaps the easiest to use because it has a user-friendly web interface (https://bcdatabaser.molecular.eco/), although it's important to note that reference databases created through this portal are publicly visible to others, so if a private database is desired then the command line tool must be used. For studies narrowly focused on fish, Meta-fish-lib (Collins et al., 2021) is specifically tailored to the creation of reference databases for this group, and it is aware of taxonomic synonyms. The refdb package (Keck & Altermatt, 2023) is convenient for users that are comfortable working in an R environment, as it provides many graphical tools for the interactive exploration and curation of the reference database. Perhaps the most sophisticated of these tools is CRABS (G.-J. Jeunen et al., 2023), which includes a variety of tools for downloading records from different sources, performing *in silico* PCR, filtering/cleaning the data, visualizing and exploring the data, and exporting the records to a variety of useful formats.

Table 9: Comparison of several reference database creation tools

Software tool	**Specific to marker?**	**Specific to taxonomic group?**	**Includes curation tools?**	**Reference**
BCdatabaser	No	No	Yes	(Keller et al., 2020)
CRABS	No	No	Yes	(G.-J. Jeunen et al., 2023)
Meta-fish-lib	Yes – but many	Yes – fish	Yes	(Collins et al., 2021)
mkCOInr	Yes – COI	No	No	(Meglécz, 2022)
Refdb	No	No	Yes	(Keck & Altermatt, 2023)
RESCRIPt	No	No	Yes	(Robeson et al., 2021)

3.6 Reference library completeness

There may be more than a billion species on Earth, of which 1.5 million have been named (Larsen et al., 2017). In comparison, the Barcode of Life Data System (Ratnasingham & Hebert, 2007), the largest dedicated collection of reference DNA barcodes, represents only 350,000 species as of mid-2023. Fortunately, the DNA barcode reference library grows at a pace of ~50% annually so the ability to give species-level identifications to environmental DNA sequences is improving every day (Porter and Hajibabaei 2018). Indeed, in an internal study performed by CEGA, when marine eDNA data were re-analyzed 18 months after an initial analysis, it was possible to increase the number of species-level assignments by 30%. This finding raises two important recommendations:

1. Unlike conventional methods of biodiversity assessment, eDNA data is not static. Indeed, data can *and should* be periodically re-analyzed to take advantage of the continuously growing reference libraries.
2. When performing comparisons between datasets—especially from one year to the next—it is necessary that *all data* be re-analyzed using a recent reference database to ensure that the

comparisons are truly apples-to-apples. Otherwise, strange artifacts are possible. For example, the number of species present in an environment may appear to increase over time simply because the number of species-level matches that can be assigned has improved as time progresses.

In the following sections, we examine the reference library completeness for several sites around the world of high O&G activity. We have focused on the marine environment because the reference libraries for terrestrial and freshwater environments have fewer gaps.

Shapefiles for each marine region were downloaded from marineregions.org, then simplified using an area-weighted Visvalingam-Whyatt algorithm in mapshaper.org. These coordinates where then uploaded to the GBIF website and used to download species occurrence data as of July, 2023 (Telenius, 2011). Finally, species records were cross-referenced with the GenBank nucleotide database (downloaded May, 2023) as a rough estimate of reference library completeness. These were further cross-referenced to the IUCN status to identify species that are deemed “near threatened”, “vulnerable”, “endangered”, and “critically endangered”. It should be noted that even though it’s the most comprehensive resource available, the data in GBIF is not exhaustive and has its own biases. For example, according to GBIF the number of species known to occur in the North Sea is greater than the number of species in the South China Sea—which is almost certainly false but is an artifact of the relative intensity of biodiversity research that occurs in these various regions.

3.6.1 North Sea

GBIF lists 36.0K known species within the North Sea[3], as defined by a shapefile obtained from marineregions.org (Figure 5). Of these, approximately 77% are present in the GenBank nucleotide database (Figure 6).

Figure 5: Map of the area of the North Sea used to query species occurrences.

[3] GBIF.org (23 July 2023) GBIF Occurrence download https://doi.org/10.15468/dl.nmc44e

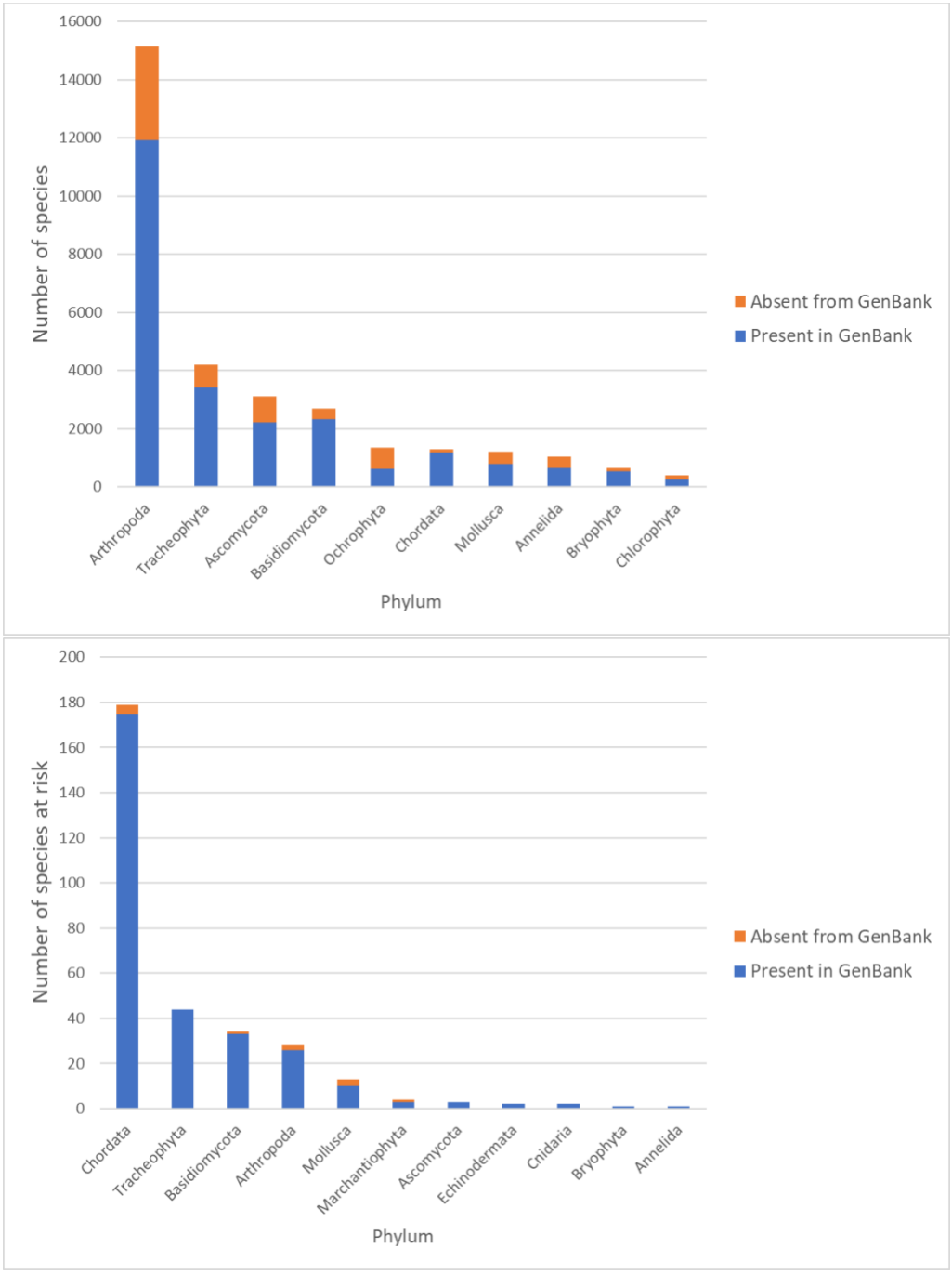


Figure 6: Approximately 77% of known species across the top ten phyla in the North Sea are present in GenBank (top figure). For species at risk, the reference library is 96.5% complete (bottom figure).

3.6.2 Gulf of Mexico

GBIF lists 27.8K known species within the Gulf of Mexico[4], as defined by a shapefile obtained from marineregions.org (Figure 7). Of these, approximately 65% are present in the GenBank nucleotide database (Figure 8).

Figure 7: Map of the area of the Gulf of Mexico used to query species occurrences.

[4] GBIF.org (23 July 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.3eyy5a

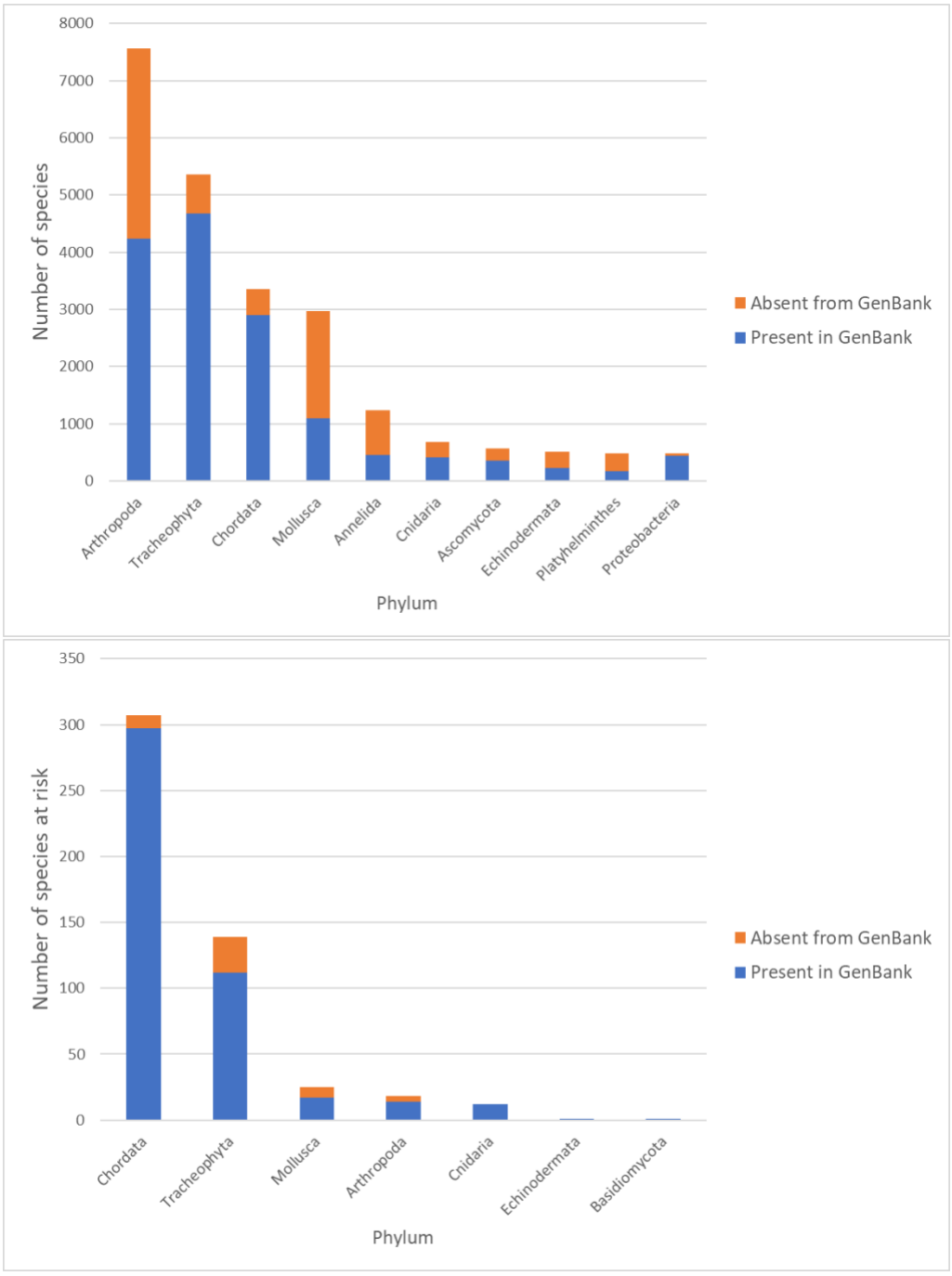


Figure 8: Approximately 65% of known species across the top ten phyla in the Gulf of Mexico are present in GenBank (top figure). For species at risk, the reference library is 90% complete (bottom figure).

3.6.3 Arabian Sea

GBIF lists 9.9K known species within the Arabian Sea[5], as defined by a shapefile obtained from marineregions.org (Figure 9). Of these, approximately 73% are present in the GenBank nucleotide database (Figure 10).

Figure 9: Map of the area of the Arabian Sea used to query species occurrences.

[5] GBIF.org (23 July 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.3qden7

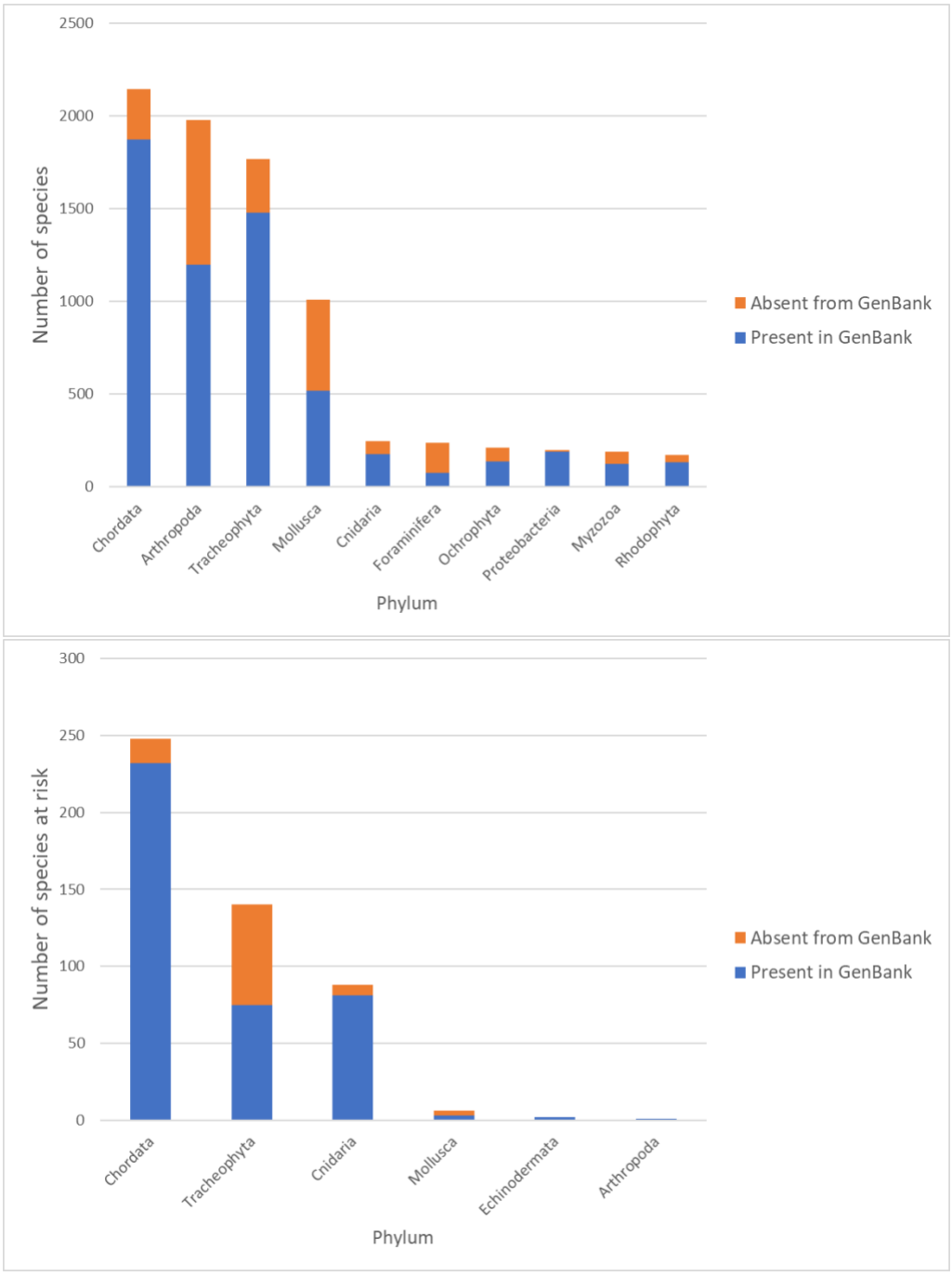


Figure 10: Approximately 73% of known species across the top ten phyla in the Arabian Sea are present in GenBank (top figure). For species at risk, the reference library is 81.2% complete (bottom figure).

3.6.4 South China Sea

GBIF lists 27.1K known species within the South China Sea[6], as defined by a shapefile obtained from marineregions.org (Figure 11). Of these, approximately 72% are present in the GenBank nucleotide database (Figure 12).

Figure 11: Map of the area of the South China Sea used to query species occurrences.

[6] GBIF.org (23 July 2023) GBIF Occurrence Download https://doi.org/10.15468/dl.rhtpc5

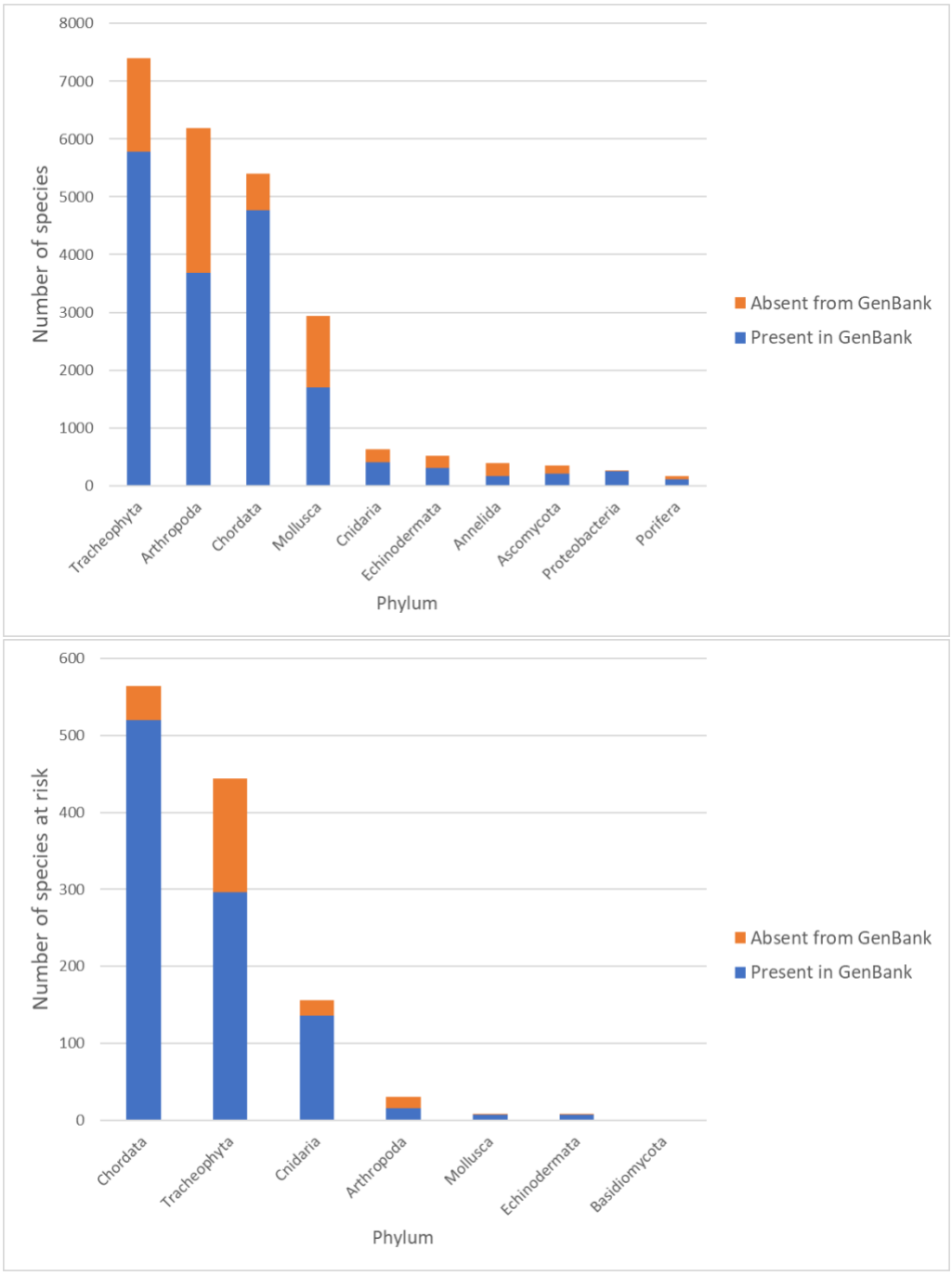


Figure 12: Approximately 72% of known species across the top ten phyla in the South China Sea are present in GenBank (top figure). For species at risk, the reference library is 81.1% complete (bottom figure).

4 Data analysis and interpretation

4.1 Interpretation

4.1.1 False Positive and False Negative Detections

An important aspect to interpreting metabarcoding data, as with any other biodiversity dataset, is understanding the sources and frequencies of false positives and negatives to mitigate their occurrences and use the data appropriately. Several practices must be applied rigorously in field sampling and laboratory processing to reduce the likelihood of either false positives or negatives, but here we focus on steps that should be taken during bioinformatics processing, data analysis, and interpretation. Despite the application of rigorous protocols that reduce the likelihood of false positives and false negatives, they cannot be eliminated entirely thus, the analysis and interpretation of metabarcoding data should account for this possibility.

4.1.1.1 Sample-level False Positives

We define false positives as detections of taxa or sequence variants in samples where that taxon's DNA or that sequence variant were not present at the point and/or time of collection (Darling et al., 2021; Drake et al., 2022). A different type of false positive can arise when a taxon's DNA is present at the point and/or time of collection, but the taxon is not present at that point or time (Darling et al., 2021; Jerde, 2021). This type of false positive results from environmental rather methodological factors and is discussed in the section *Site-level Errors Arising from Environmental Conditions* below. This section focuses on where and how false positives arise during the metabarcoding workflow and how to account these, where possible. False positives can arise during field sampling through external contamination such as improper decontamination of sampling equipment (Burian et al., 2021). False positives can also occur in the lab and bioinformatics workflows through external contamination (e.g., DNA in reagents), cross-contamination between samples, index hopping or tag jumping, artefacts (i.e., PCR or sequencing errors), chimeras, and incorrect taxonomic identifications (e.g., due to database errors, pseudogenes) (Bell et al., 2019; Burian et al., 2021; Drake et al., 2022; Ficetola et al., 2015; Graham et al., 2021; Santoferrara, 2019).

Many sources of false positives can be controlled or reduced through the bioinformatics process. Below, we note the bioinformatics decisions and steps that can be taken to reduce the incidence of false positives and refer users to any other sections of this document where these steps have been previously discussed. Any steps that have not been discussed in other sections of this report are described in more detail here.

Several mitigation steps that can be applied during bioinformatics have been discussed in other sections: denoising to remove artefacts (Section 2.5), chimera removal (Section 2.6), and robust procedures for taxonomic assignment (Section 2.6). Mitigation measures that have not been covered in previous sections include: decontamination based on negative controls and minimum read thresholds (Alberdi et al., 2018; Drake et al., 2022). Decontamination approaches rely on rigorous field/lab protocols that include negative controls throughout the workflow (e.g., field, extraction, PCR negative controls). There are several approaches to use detections in negative controls to mitigate contamination in samples, including removal of all ASV/OTUs detected in negative controls from samples (Drake et al., 2022; Karstens et al., 2019), subtracting raw reads recovered from ASV/OTUs in negative controls from associated samples (Andruszkiewicz et al., 2017; Bell et al., 2019), using a relative abundance-based

subtraction approach (e.g., microDecon package (McKnight et al., n.d.)), prevalence-based approaches (e.g., decontam package (Davis et al., 2018)), frequency-based approach using DNA concentrations (e.g. decontam package (Davis et al., 2018)), and predicting contamination based on known environments/contaminants using source tracking (Karstens et al., 2019; Knights et al., 2011). There is not a broad consensus on the choice of decontamination approach to use, but rather the choice of appropriate decontamination approach depends on project goals and the contamination data observed for a given sample set (Karstens et al., 2019). Since the optimal approach is project-dependent, we recommend that whatever method is used for decontamination, it be tracked and reported as appropriate.

Minimum read thresholds can be applied to remove artefacts, index hops, and cross-contamination. The rate of cross-contamination, index hopping, and artefacts depend on a laboratory practice and parameters (e.g., increasing sequencing depth increases the artefact abundance (Alberdi et al., 2018; Ficetola et al., 2016)) and thus have led to a range of approaches. Thresholds can be applied equally across all samples and/or sequence variants (e.g., no filter (Lacoursière-Roussel et al., 2018), singleton removal (Bylemans et al., 2019; Guardiola et al., 2016), or removal of low abundance variants in a sample with a threshold ranging from 3-1000 (Cowart et al., 2015; Drake et al., 2022; Wangensteen et al., 2018)). Alternatively, they can be applied on a proportional basis in different contexts. Reads below a certain proportion of the total reads across all OTU/ASVs and samples can be removed (Braukmann et al., 2019; Klymus et al., 2017). Reads in a sample below a certain proportion of the total reads in the sample can be removed (Lopes et al., 2017; McInnes et al., 2017; Yamamoto et al., 2017). Reads in a sample with an abundance less than a proportion of the total OTU/ASV read count across all samples can be removed (Lopes et al., 2017; Pont et al., 2018; Wangensteen et al., 2018). Additional measures include removal of OTUs/ASVs that don't have reads across multiple technical replicates (i.e., PCR replicates (Laroche et al., 2017; Lim et al., 2016)) or markers (González et al., 2023). This is not an exhaustive list of approaches to apply minimum read thresholds and there is not a broad consensus on the choice of minimum read approach and threshold to use, but rather the choice of appropriate thresholds should be optimized depending on project goals, reads recovered in positive and negative controls, and unassigned indexes/tags for a given sample set/sequencing run (Alberdi et al., 2018; Drake et al., 2022; González et al., 2023; Karstens et al., 2019). Since the optimal approach is project-dependent, we recommend that whatever method is used for decontamination, it be tracked and reported as appropriate.

4.1.1.2 Sample-level False Negatives

False negatives can arise in several different ways and can thus be defined in several ways. We define two types of false negatives that are relevant to the bioinformatics processing. First, a sequence variant was present in a sample, but the sequence variant was not present in the data output after bioinformatics processing (Doi et al., 2019; Ficetola et al., 2015; McClenaghan, Compson, et al., 2020; Zinger et al., 2019). Second, a taxon's DNA was present in the sample, but the taxon was not identified from the sequence data (Schenekar et al., 2020). A third type of a false negative can arise at the sample level when a sequence variant was present in the environment at the time and location of sampling but was not captured in a sample (Burian et al., 2021; Doi et al., 2019; Ficetola et al., 2015; McClenaghan, Compson, et al., 2020). Mitigation measures to minimize this type of false negative must be applied during sampling design and sample collection and won't be discussed here. This type of false negative is however relevant to the discussion in the section *Interpreting Sample-level Errors*. Finally, false negatives

can also occur at the site-level from environmental conditions, rather than due to the genomics workflow. These are discussed in the section *Site-level Errors Arising from Environmental Conditions* below.

The two false negatives defined above that are relevant to bioinformatics may occur due to sample degradation, low sensitivity of PCR assay or PCR assay bias, DNA extraction stochasticity, PCR stochasticity, PCR inhibition, low sequencing depth, bioinformatic filtering is too stringent (e.g., minimum sequence copy number threshold), true sequence variants lumped into clusters, a lack of reference sequences, or a lack of resolution between sister taxa (Alberdi et al., 2018, 2019; Burian et al., 2021; Santoferrara, 2019; Schenekar et al., 2020).

Most of these sources of false negatives need to be mitigated during the sampling design, sample collection and/or lab processing phases. There are two sources of false negatives that can be mitigated during bioinformatics: the stringency of bioinformatic filtering and the lumping of true sequence variants during denoising or OTU clustering. Parameter selection and optimization for denoising and clustering are covered in Section 2.5. The stringency of bioinformatic filtering refers back to the methods for decontamination and minimum read thresholds used to reduce false positives. Strict abundance-filtering methods can introduce false negatives and it has even been shown that these steps can introduce false negatives at higher rates than they eliminate false positives therefore they must be applied carefully (Littleford-Colquhoun et al., 2022). Enforcing multiple marker or multiple replicate detection thresholds also increase false negative rate considerably (N. T. Evans et al., 2017). Application of bioinformatics filters must consider project-specific objectives and, as noted previously, we recommend that whatever bioinformatic filtering approaches are used, they be tracked and reported as appropriate. See the section *Balancing False Positive & False Negatives* below for a discussion of how favouring fewer false positives or fewer false negatives may affect the resulting data.

4.1.1.3 Balancing False Positive & False Negatives

Every mitigation measure applied during bioinformatics needs to be applied while considering the balance between false positive and false negative detection. A strict threshold/method will remove false positives but introduce false negative, while a less stringent threshold will reduce false negatives but retain false positives (Alberdi et al., 2018; Drake et al., 2022; Littleford-Colquhoun et al., 2022). In other words, a higher stringency will provide high certainty but less sensitivity. Often, minimizing false positives is prioritized over false negatives, because the absence of DNA does not prove the absence of a taxon, while the presence of DNA is usually interpreted as proving the (potentially erroneous) presence of a taxon. However, the tolerance for false positives and false negatives will vary depending on project objectives and therefore, the choice of bioinformatics parameters and thresholds will be selected on a project specific basis. For example, a project looking for an invasive or endangered species may opt to use less stringent threshold(s) to obtain a lower false negative rate and tolerate a higher risk of false positives. False positive rates are proportionately higher for rare taxa (base rate fallacy (Darling et al., 2021)) therefore a tolerance for false positives is important when looking for a rare target. Detections could be followed up with further surveys for confirmation. A project looking at ecosystem-level impacts may opt for more stringent measures to reduce false positives at the risk of a higher false negative rate. This approach will likely not impact community-level trends observed with metabarcoding data and thus allow inference about ecosystem level impacts (Wilding et al. 2023; Porter et al. 2019). Whatever

bioinformatic filtering methods are applied, practitioners should be able to justify their choice of parameters and thresholds in the context of that project or program's objectives.

4.1.1.4 Interpreting Sample-level Errors

While sample-level errors may persist in a dataset despite employing the best available methods to minimize them, thoughtful interpretation can mitigate the impacts these errors have on conclusions made from the data. Several of these approaches are also used with other sampling methods to account for errors that arise from those approaches.

It is important to maintain consistent protocols within a monitoring program or project, so that any biases/errors that are not accounted for are consistent across the study and do not create confounding factors for sample comparisons. For example, if samples are collected over several years and each year of samples is analyzed as they are collected, taxonomy would be assigned based on different reference sequence data available each year (Schenekar et al., 2020; Taberlet et al., 2018b). To enable more robust comparisons across years, previous years' data should be re-assigned taxonomy using the latest reference database each year (see Section 3.6). All other parameters and filtering thresholds applied throughout the bioinformatics workflow should also be consistent.

The molecular assay used may be a source of false negatives due to primer bias and/or inefficient amplification. This can be mitigated by careful choice of primers and the use of multiple assays during the sampling design phase, but incorporating an understanding of the limitations and biases of whatever assays are used into the interpretation of results will generate more robust conclusions (McClenaghan, Fahner, et al., 2020). If known biases for a given assay are not mitigated through the use of multiple complementary assays during the sampling design phase, these biases should be acknowledged during interpretation.

Analytical tools designed to account for false positives and negatives and generate more robust estimates of species/taxon occupancy are available and used often for other biodiversity monitoring methods (Hamer et al., 2021; Mills et al., 2019) and, increasingly, for environmental genomics methods (McClenaghan et al. 2020; Bush et al. 2020; Pukk et al. 2021). A hierarchical occupancy modelling framework can be applied to account for false negatives at multiple scales (i.e., during sample collection and during PCR amplification) for single species and multi-species data (McClenaghan, Compson, et al., 2020; Schmidt et al., 2013). Hierarchical frameworks for the inclusion of false positives are available for single-species models although they generally require some prior knowledge or complementary methods (Guillera-Arroita et al., 2017). Multi-species models accounting for false positives and false negatives are in development but are more difficult to apply when *a priori* information is required (Burian et al., 2021). Hierarchical occupancy models are especially useful when paired with environmental or methodological data that varied across samples and that is relevant to species distributions and/or probabilities of detection. This framework accounts for variability in the probability of detection (i.e., rate of false positives and negatives) across sites and samples when making conclusions about species distributions. When the appropriate data is available, we recommend using an analytical approach such as this to account for imperfect detection in metabarcoding data. These models are also discussed in the section *Community Analyses*.

Given the many definitions of false positive and false negative that exist and the contexts in which they can apply, we recommend that explicit definitions be provided when using these terms or use more specific terms to indicate how and where in the workflow these errors arise.

4.1.1.5 Site-level Errors Arising from Environmental Conditions

False positives or false negatives may occur at the level of sampling sites due to environmental conditions. In this context, a false positive is the presence of a taxon's DNA at a sampling location and time, when the taxon was not present at that sampling location and time. A false negative is the absence of a taxon's DNA at a sampling location and time when the taxon was present at that sampling location and time. These types of errors are the result of environmental and biological conditions, not the genomics methods used. Factors contributing to site-level errors include DNA persistence and degradation, organism physiology (i.e., eDNA production rate), hydrological conditions, and environmental conditions such as temperature or pH (Burian et al., 2021).

An understanding of local hydrology may inform the interpretation of results and can be integrated into models of distribution/abundance when this data is available (Burian et al., 2021; Carraro et al., 2020; Fremier et al., 2019). This is likely more achievable for small lentic or lotic freshwater systems, but in large systems and marine environments the hydrology can be quite complex and/or hard to measure (e.g., deep ocean currents). For terrestrial environments, including both soil and air substrates, the movement of eDNA is not well understood (Taberlet et al., 2018c). Since information is lacking on eDNA transport in the environment, there are limited mitigation options beyond gathering more project-specific data. Users should be aware of the possibility of eDNA movement impacting detection patterns and incorporate this into sampling design and interpretation where relevant. Similarly, knowledge of DNA production, persistence, and degradation rates for the organisms, environments, and substrates of interest can inform the interpretation of eDNA results. This is an active area of research (Barnes et al., 2014a; Barnes & Turner, 2016; Collins et al., 2018; Dejean et al., 2011; Foucher et al., 2020; Nielsen et al., 2007; S. A. Wood et al., 2020), however with the wide range of organisms captured in eDNA surveys and the variety of environmental mediums and conditions spanned by eDNA studies, there is still limited information available. When information is available for a study organism or environment, it should be included in the sampling design and interpretation of eDNA results, but where information is lacking on eDNA production, persistence, and degradation, users should be aware of the possibility of these factors impacting detection patterns and interacting with eDNA movement.

There is also a possibility of sampling-independent contamination (e.g., ballast water, predator feces, human activities) which can lead to false positive errors (Burian et al., 2021; N. T. Evans et al., 2017). These sources of false positive are harder to mitigate so users should remain aware of the possibility and probability of such events occurring in the study area. Careful sampling design with repeated sampling time points paired with bioinformatics filtering thresholds requiring detections across multiple replicates may mitigate the risks of this type of false positive but would need to be assessed on a project-specific basis.

4.1.2 Noise

Metabarcoding data, like all biodiversity datasets, includes noise. Some of this noise arises from false positives, but much of it arises from stochasticity between replicates throughout the workflow (e.g., biological replicates during sampling, subsampling of DNA extracts for PCR replicates) (Wilding et al.,

2023). Pseudogenes can be an additional source of noise, where many sequence variants can arise from a single individual. These may all be correctly identified and not represent false positives but artificially inflate the read counts and number of unique sequences from certain individuals, thus adding noise to the data (Graham et al., 2021). There exist methods that can attempt to identify and remove pseudogenes using machine learning, although this must be trained in a marker-specific manner and can only identify pseudogenes that have accumulated enough mutations to look distinct from real genes (Porter & Hajibabaei, 2021). Having too much noise in a dataset can mask true biodiversity patterns either at the level of individual taxa or the whole-community (Graham et al., 2021; Wilding et al., 2023).

Robust sampling design, sample collection, and lab processing methods will all reduce noise in metabarcoding data. Mitigation measures to reduce noise that can be applied during bioinformatics and analysis have already been discussed in the section *Sample-level False Positives* above. In order to maximize the signal-to-noise ratio, particularly for whole community assessments along impact gradients, quite stringent thresholds to remove low frequency or low abundance ASVs may be applied (Wilding et al., 2023). Noise is non-random and can be accounted in modeling approaches (Gold, Shelton, et al., 2023). Occupancy modeling provides a means to account for some of the stochasticity inherent in genomics workflows however modeling approaches specifically designed to account for noise remain a frontier for development.

4.1.3 Quantitative vs Presence/Absence

The use of eDNA data for compositional or quantitative analyses is debated (Deagle et al., 2013; Di Muri et al., 2020; Goldberg et al., 2016; Piñol et al., 2018; Shelton et al., 2022). eDNA data generates counts of unique DNA sequences (or taxa) in a given sample, which could potentially be used in quantitative analyses. However, there are biases that can be introduced into these count data through the laboratory process (especially differential binding of PCR primers to environmental DNA from different taxa), such that the read counts in the sequencing data do not accurately reflect the amount of DNA in the environmental sample (Shelton et al., 2022). Additionally, there are biological and environmental factors that can affect the amount of DNA from a taxon present in the environment, meaning the amount of DNA in a sample does not necessarily reflect the abundance or biomass of that organism in the environment (Goldberg et al., 2016). There are many factors that can influence each of these two processes. For example, the metabolic rate or age of an organism can influence how much DNA is released into the environment (Lacoursière-Roussel et al., 2016; Rourke et al., 2022; Takeuchi et al., 2019). Temperature and microbial activity can affect how quickly DNA degrades, thus affecting the amount of DNA from an organism that persists in the environment (Barnes et al., 2014b; Joseph et al., 2022; Rourke et al., 2022). In the lab, the efficiency of each primer set varies across taxonomic groups, therefore any general primer set will preferentially amplify certain taxonomic groups over others (Elbrecht & Leese, 2015; Piñol et al., 2018; Rourke et al., 2022). Despite the technical, biological, and environmental factors that impact read counts, many studies show strong correlations between sequence read counts in metabarcoding data (raw and transformed) and organism biomass in the environment (Di Muri et al., 2020; Ershova et al., 2021; N. T. Evans et al., 2016; W. Li et al., 2021; Skelton et al., 2022; Thomas et al., 2016; Tsuji et al., 2022) and read counts are often used for ecological analyses (Keeley et al., 2018; Marquardt et al., 2016; Mauffrey et al., 2021; Numberger et al., 2019; Ratcliffe et al., 2021). Relative abundances can also be derived from metabarcoding data through presence/absence across high-replicate samples using occupancy modeling (Bush et al., 2023). This procedure assumes that

the likelihood of detecting a species in a given sample is directly proportional to its abundance. For example, a species that is detected in 40 out of 50 samples is believed to have a higher relative abundance than a species that was only detected in 10 out of 50 samples.

If sequence read counts will be used for quantitative analyses, a robust laboratory workflow designed to reduce and/or account for biases should be used where possible (Luo et al., 2023). Depending on the study goals, different analytical approaches may be used to account for biases. For example, applications comparing within-species abundance across samples may take a different analytical approach that those comparing within-sample across species abundance (Luo et al., 2023; Shelton et al., 2022). Analyzing and interpreting quantitative results will depend on the workflow employed in the laboratory and the application of the data. We provide some examples below; however, this is a rapidly evolving area of research and methods will be subject to change as new approaches emerge.

To reduce primer bias during amplification and generate more accurate sequence read counts, primer sets that perform better for this purpose can be selected. Universal primer sets that show less biased results—and therefore more accurate quantitative results—should be used (Piñol et al., 2018). Primer sets can be also designed to generate less biased results for quantitative applications (e.g., the LERAY-XT degenerate COI primer set (Ershova et al., 2021)). The choice of target gene region can also improve quantitative results. For example, the photosynthetic gene psbO in phytoplankton is exclusively present in photosynthetic organisms and exists primarily in one copy per genome, reducing the bias that can be introduced when gene copy number varies between organisms (Pierella Karlusich et al., 2023). Marker selection for quantification must be done at the design phase. After laboratory analysis there are several analytical approaches that can be applied to account for primer bias and enable within sample across species comparisons. The most common approach is the application of correction factors. These correction factors may be derived from the results of mock community analysis (Krehenwinkel et al., 2017; Matesanz et al., 2019; Thomas et al., 2016), from allometric scaling (Yates et al., 2022), or from the number of gene copy numbers across species (J. L. Martin et al., 2022). The use of these correction factors improve the correlation between sequence read counts and the original DNA abundance in a sample (Krehenwinkel et al., 2017) or the abundance of the organism in the environment (Yates et al., 2022). Alternatively, emerging model-based approaches directly model amplification efficiency across taxa to correct data and generate accurate estimates of community composition (Shelton et al., 2022).

The previously described methods aim to reduce biases across taxa within a sample. To enable robust comparisons within a taxon across samples, the noise that is generated between samples must be accounted for. The use of internal standards or spike-ins is the most common approach to account and correct for this type of noise (Luo et al., 2023; Smets et al., 2016; Tsuji et al., 2020). Other studies use qPCR or ddPCR based quantification of total eDNA in a sample to correct for variation in starting eDNA concentrations (Pont et al., 2023; Van Bleijswijk et al., 2020). qPCR and ddPCR can accurately quantify the amount of DNA in a sample enabling an estimate of the amount of eDNA for each organism in each sample when paired with compositional metabarcoding data (Zemb et al., 2020). However, as noted above, compositional data within a sample can be impacted by primer biases and the relationship between an organism's abundance/biomass and the amount of eDNA in the environment is affected by several environmental and biological conditions. These approaches to reduce noise that is generated between samples must be incorporated during the laboratory analysis steps and also require analytical correction.

The detection/non-detection of a species across multiple samples can be used as an index to compare within species across sample sets abundance. There is generally a positive relationship between species occurrence and abundance (Gaston et al., 2000), meaning that with sufficient sampling to accurately estimate a species' occurrence in a given area, the probability of occurrence could be used as a measure of relative abundance (i.e., areas with a higher probability of occupancy have a higher relative abundance than areas with a lower probability of occupancy) (MacKenzie & Nichols, 2004). This is sometimes referred to as a semi-quantitative measure of abundance. This method reduces, but does not remove, noise introduced during the metabarcoding workflow and is not recommended to compare across species (Luo et al., 2023; Sard et al., 2019; J. Yang et al., 2017).

No matter what approach is used, reporting should be clear and specific about the approach that was implemented, what the potential sources of error are, and it should be made clear that eDNA data are the result of the abundance of an organism's DNA in the environment. Without knowledge of species-specific biology and environmental conditions, extrapolating these results to estimate absolute abundance or biomass of the organism will include additional sources of uncertainty. eDNA data generally performs better at estimating biomass than abundance (Elbrecht & Leese, 2015; Lamb et al., 2019). Most traditional indices are based on abundance not biomass, therefore even where accurate quantitative estimates can be made, metabarcoding data may not be compatible with traditional indices. Some studies show consistent results across metabarcoding and morphological datasets for certain indices (e.g., AMBI (Aylagas et al., 2018)). Additionally, new indices are emerging that are designed for this type of data (Mächler et al., 2021). Further discussion of this can be found in the section *Ecological Indices.*

4.1.4 Controlling for Sampling Effort

It is well established that uneven sampling effort can bias biodiversity inferences across samples or sites. This is not unique to environmental genomics approaches and several resources are available that discuss this (R. Colwell & Coddington, 1994; Gotelli & Colwell, 2001; Moreno & Halffter, 2000) as well as methods to reduce or address these biases (Buddle et al., 2005; Oliveira et al., 2017; Pardo et al., 2013; Stolar & Nielsen, 2015). These resources focus on non-molecular methods; however, laboratory analysis of environmental genomics samples creates additional opportunities for uneven sampling effort to be introduced into the biodiversity characterization process. There are several approaches to mitigate this. This section describes the sources of biased sampling effort during laboratory processing and potential methods for controlling for uneven laboratory sampling effort that occurs with environmental genomics samples. The same approach may not apply across all projects and scenarios thus these are presented as options and users should be aware of how sampling effort applies to environmental genomics samples to evaluate various approaches that may be used.

The most common source of variation in sampling effort during laboratory processing is sequencing depth, also referred to as depth of coverage per sample, or library size (McMurdie & Holmes, 2014). Sequencing libraries are generally prepared by combining samples in equimolar concentration however, due to random sampling during sequencing the resulting sequencing depths across samples can vary by orders of magnitude (McMurdie & Holmes, 2014). Samples may also be sequenced on multiple different sequencing runs further contributing to variation in sequencing depth between samples. Higher sequencing depths yield more unique sequences and more taxa (Singer et al., 2019), and thus impact alpha and beta diversity estimates (Shirazi et al., 2021). Rare or low abundance taxa are most influenced

by variation in sampling effort (Shirazi et al., 2021). Another factor contributing to laboratory sampling effort is variation in replication and pooling at different stages (e.g., DNA extraction, PCR). Replicates from a given step can be added and carried through the entire process and subsequently pooled back together. Typically, protocols are applied consistently across samples from a sample set, but in some cases if samples are processed at different times (e.g., time series sampling) methods may have evolved resulting in different sampling effort across samples. Wherever sampling effort differs between samples, this should be acknowledged and/or accounted for in analysis and interpretation.

To address uneven sequencing depth, normalization of read counts based on sequencing depth can be done using a variety of scaling factors, the simplest of which divides each taxonomic unit's read count by total sequencing depth for that DNA marker, can be used to generate relative frequency data (McMurdie & Holmes, 2014; Taberlet et al., 2018a). This simple approach can overcome biases as a result of sampling effort in some cases but is not appropriate for data being used to estimate or compare richness among samples and yields biased differential abundance estimates (Bullard et al., 2010; McMurdie & Holmes, 2014; Weiss et al., 2017). Despite these limitations, this approach is widely used (Muha et al., 2021; Schenk et al., 2019; Skidmore et al., 2022). Other scaling factors used include cumulative sum scaling, trimmed mean of M component scaling, and, more recently, Analysis of Compositions of Microbiome with Bias Correction (ANCOM-BC), which performed the best in a comparison of common scaling factors (Lin & Peddada, 2020).

Rarefaction curves provide another approach to control for uneven sampling effort. Rarefaction curves plot taxon recovery (e.g., # species, # OTUs) against sampling effort (e.g., sequencing depth, number of replicates) to visualize whether a sample has reached saturation; that is, almost all taxa that could be detected, even with increased sampling effort, were detected (Matthews et al., 2021; Shirazi et al., 2021). If all samples have reached saturation, variation in sampling effort should not have a strong impact on diversity estimates (Taberlet et al., 2018a). If samples have not reached saturation, samples that did not reach saturation can be discarded or the sampling effort can be standardized by rarefying the data set (i.e., randomly subsampling to a given sampling effort) (Taberlet et al., 2018a). Both approaches result in a loss of information and reduced statistical power since data are being discarded. For example, when sequencing depth is standardized using rarefaction, the uncertainties in taxon relative abundances across samples increases which then results in less statistical power to detect differences between groups or samples (McMurdie & Holmes, 2014). When rarefying data, the sampling effort used as the subsample size should be selected to minimize the number of samples or amount of data excluded from the analysis and maximize the proportion of the sample diversity retained. It has been suggested that samples should be rarefied to an equal diversity coverage (e.g., 95%) per sample based on rarefaction curves, which would result in samples being rarefied to different levels sampling effort depending on site or sample specific characteristics (Taberlet et al., 2018a). However, determining the optimal rarefaction threshold for new empirical data may not always be possible (McMurdie & Holmes, 2014). Despite these limitations, this approach is widely used (Lejzerowicz et al., 2021; Mächler et al., 2021; Matthews et al., 2021).

For differences in sampling effort at the level of replication, extrapolation can generate estimates of species richness that account for sampling effort (Taberlet et al., 2018a). For example, the Chao1 index estimates species richness at a site using the variation in diversity recovered between replicates (Chao & Chiu, 2016). These methods are sensitive to rare taxa and may be impacted by artefacts that arise in

metabarcoding data, therefore they may not perform well in accounting for variation in sequencing depth.

There are several modelling approaches that can be used to make inferences about species distributions and community composition while accounting for variation in sampling effort across samples and sites. A hierarchical occupancy modeling framework can include sequencing depth as a covariate at the level of replicate samples and can accommodate varying levels of replication across multiple sampling levels (e.g., sites, samples, PCR replicates) (McClenaghan, Compson, et al., 2020; Willoughby et al., 2016). Other modeling frameworks at their current stage of development may be able to incorporate sampling effort under sampling design parameters (e.g., HMSC (Tikhonov et al., 2020)), but do not account for the possibility of imperfect detection that occupancy models include.

Since many different approaches are available and the optimal approach depends on project specific goals, the method used to account for variation in sampling effort should be reported, with justification to support the analyses being conducted for a given project.

4.2 Bioindicators & Biotic Indices

Ecosystems are complex systems to monitor with a large number of interacting biotic and abiotic factors. Bioindicators were developed as a simplified approach to monitor environmental conditions, ecological processes, and/or biodiversity. Bioindicators are selected as representative or aggregated responses for the ecosystem (Holt & Miller, 2010). Sample and data collection efforts can then be focused on bioindicators identified for the type of stress or disturbance in the environment of interest (Holt & Miller, 2010). Environmental genomics can be used to generate data on known bioindicators or be used to identify new bioindicator taxa since environmental genomics tools generate data on whole communities, including taxa difficult to capture or identify with traditional methods (He et al., 2020). Data on bioindicator taxa are often used to calculate biotic indices (Lanzén, Dahlgren, et al., 2021). Biotic indices integrate information from multiple taxa, most often bioindicator taxa, into a single value that provides a metric for ecosystem health or quality (Borja et al., 2000). Biotic indices are widely used and accepted by regulatory agencies for biomonitoring (Monaghan & Soares, 2012). Environmental genomics data can be used to calculate established biotic indices or used to develop new biotic indices (Lanzén, Dahlgren, et al., 2021). The use of environmental genomics data for bioindicator and biotic index analyses is quickly evolving with new research and efforts to ensure eDNA-based biodiversity data can integrate into monitoring frameworks. Environmental genomics data have unique features compared to typical morpho-taxonomic data, so there are considerations and limitations that users should be aware of when interpreting the resulting indices.

4.2.1 Bioindicators

Environmental genomics can be used to monitor previously established indicator species or groups (Capurso et al., 2023; Carew et al., 2013; Hajibabaei et al., 2011; He et al., 2020) and to classify ecosystem status based on these bioindicators (He et al., 2020; Kuntke et al., 2020; Miyata et al., 2022). Alternatively, environmental genomics data can be used to identify new bioindicator taxa if sampling is conducted along an appropriate gradient (Laroche et al., 2016; Pawlowski et al., 2016). Indicator analyses can also be used to evaluate community-level biodiversity patterns and identify the species driving differences in biological communities (G. Jeunen et al., 2019; Krah & March-Salas, 2022)(see more in *Community Analyses* section). Using environmental genomics data to identify bioindicators

broadens the scope of biomonitoring by including taxonomic groups that were not previously used due to difficulty in collection or identification (Pawlowski et al., 2018). When using environmental genomics data to identify new bioindicators, the bioindicators may be identified using taxonomic information (e.g., species names) or unique sequences (e.g., OTUs) which allow a larger portion of the metabarcoding data to be used (Pawlowski et al., 2016; Stoeck, Kochems, et al., 2018).

Methods for the classification of ecosystems status and identification of indicator taxa based on environmental genomics data are generally consistent with the methods used for morpho-taxonomic data, however there are some important factors to consider when implementing these methods. If using unique sequences as bioindicators, users must consider that DNA sequences from the same individual could yield different OTUs/ASVs depending on the bioinformatics parameters used during analysis. In order to achieve comparable OTUs/ASVs, the same bioinformatics pipeline must be followed to enable cross-study or broad-scale use of unique sequences as indicators. Additionally, the use of multiple DNA markers can bias results at the unique sequence level if data from multiple markers are combined. Some individuals or taxa may be represented by both markers while other are not due to differences in primer binding. Individuals or taxa can be double counted giving them more weight in downstream analyses. Using taxonomic data when combining multiple markers avoids this issue, however most studies identifying new indicator taxa with metabarcoding data have used a single DNA marker (Kelly et al., 2020; Laroche et al., 2016; Pawlowski et al., 2016; Stoeck, Kochems, et al., 2018) or analyzed markers separately (Lanzén, Dahlgren, et al., 2021). Both presence/absence data and quantitative data can be used to identify indicator species and assess ecological status (Alexander et al., 2020). The biases associated with using environmental genomics data quantitatively are discussed in the *Quantitative Analyses* section (4.1.3). However, quantitative data is often used to identify indicator taxa, with relative abundance across samples being used as a quantitative measure (He et al., 2020; Lanzén, Dahlgren, et al., 2021; Laroche et al., 2016; Pawlowski et al., 2016).

4.2.2 Biotic Indices

A biotic index is a value generated from the assessment of indicator organisms, which provide information on ecological status of the environment being surveyed by comparison with presence and abundance patterns in reference conditions (Pawlowski et al., 2018). Biotic indices range in complexity from simple univariate indices (e.g., AZTI's Marine Biotic Index (AMBI) (Borja et al., 2000)) to complex, multimetric indices (e.g., Multimetric Index for Stream Acidity (MISA) (Birk et al., 2012)). The most commonly used indices, and those accepted by regulatory agencies, were developed based on morpho-taxonomic data and include some ecological and/or functional information (e.g., sensitivity to disturbance). Metabarcoding data can be used to calculate these same indices and often metabarcoding generates more data than morpho-taxonomic methods, by increasing taxonomic resolution and detecting all life stages (Pawlowski et al., 2018). However, there are limitations to the use environmental genomics data within traditional morpho-taxonomic biotic indices, which has spurned the creation of indices based on molecular data (Keeley et al., 2018). Both the integration of metabarcoding data into traditional indices and the development of new molecular indices are active areas of research. We discuss the advantages and limitations of commonly used approaches and provide recommendations for reporting to account for the range of approaches available.

4.2.2.1 Integrating Environmental Genomics into Existing Indices

There is general agreement in the results of pre-existing indices for ecological status assessment calculated using morpho-taxonomic data and metabarcoding data (e.g., the Norwegian Sensitivity Index (Lanzén, Dahlgren, et al., 2021); AMBI (Aylagas et al., 2018); Swiss Diatom Index (Visco et al., 2015); IBCH (Brantschen et al., 2021)). These indices integrate ecological information about the taxa detected with relative abundance data. As such, taxonomic assignment is a critical step in using metabarcoding data for biotic indices to link sequence data to ecological information. Typically, only a fraction of metabarcoding data gets used to calculate the indices, because a large portion of sequences are not assigned taxonomy or relevant ecological information is not available for the taxa identified (Mauffrey et al., 2021; Pawlowski et al., 2018). Furthermore, the taxa that are identified can be biased by reference database coverage, which may be more comprehensive for certain taxonomic groups compared to others (Aylagas et al., 2014; Hajibabaei et al., 2019). Appropriate DNA marker selection for the bioindicators groups associated with a given index is essential (Aylagas et al., 2014). Multiple DNA markers can be used to overcome some bias present in reference databases, however quantitative data from multiple markers should be analyzed separately (Lanzén, Dahlgren, et al., 2021; Mauffrey et al., 2021). Quantitative data (i.e., relative abundance across samples) is often used to calculate indices and has performed well for status assessment, despite known biases with quantitative data generated by metabarcoding (Aylagas et al., 2018; Mauffrey et al., 2021; Pawlowski et al., 2018; Sanchez et al., 2022). Some indices can also be calculated using presence/absence data to avoid these biases (Fernández et al., 2019), however it has been shown that quantitative data generates results more similar to morpho-taxonomic data for certain indices (Aylagas et al., 2018). There are many biological and technical factors that influence the calculation of indices using metabarcoding methods compared to morpho-taxonomic methods but given the agreement between these two approaches across multiple indices and the widespread use of these indices in regulation, we suggest that metabarcoding data can be used to calculate traditional indices where agreement between methods has been shown.

4.2.2.2 Developing New Indices

4.2.2.2.1 New Bioindicators

New biotic indices are being developed based on metabarcoding data due to the increased taxonomic breadth and ease of identification achieved for certain taxonomic groups (Aylagas et al., 2017; Pawlowski et al., 2016, 2018). Where these indices rely on taxonomic information, the limitations of these new indices are the same as those discussed above. One additional consideration is that the same reference database should be used for index development and ecological assessment with that index (Lanzén, Dahlgren, et al., 2021; Visco et al., 2015). With large gaps that exist in reference databases and new records continuously being added, an updated reference database can have a large impact on the taxonomy assigned to metabarcoding data (Hestetun et al., 2020; Morard et al., 2019). Where these new indices being developed do not rely on taxonomic information, they are discussed below in the *Taxonomy Free Indices* section.

4.2.2.2.2 Taxonomy Free Indices

Several biotic indices have been developed using unique sequence units (i.e., OTUs/ASVs) to maximize the use of metabarcoding data and overcome gaps in reference databases (Apothéloz-Perret-Gentil et al., 2017; Cordier, 2020; Cordier et al., 2017; Keeley et al., 2018; Lanzén, Mendibil, et al., 2021; Porter & Hajibabaei, 2020). This approach requires training data with unique sequences across samples from

known ecological status, to assign ecological values to unique sequences (Pawlowski et al., 2018). This can be a drawback as new sequences are needed to develop metrics for new environments and indicator groups (Lanzén, Dahlgren, et al., 2021), however there is also the potential for the new indices to be compatible with traditional indices and current regulations by using the same assessment categories (Cordier, 2020). Both correlative (Apothéloz-Perret-Gentil et al., 2017) and machine learning approaches (Cordier et al., 2017) have been used to develop these new metrics. These new methods show strong agreement with traditional indices or outperform them, as long as the DNA markers used capture bioindicator groups (Cordier et al., 2017, 2019; Keeley et al., 2018; Lanzén, Mendibil, et al., 2021). The DNA marker(s) used to develop the index must be selected appropriately for the environment and stressor/disturbance of interest and the same marker(s) must be used for subsequent assessment based on the index. Taxonomy free index approaches are calibrated on the unique qualities of environmental genomics data (e.g., different units of presence represented by eDNA vs. morphologically identified species), and users need to follow the same bioinformatics workflow to generate comparable data on unique sequences units (Cordier, 2020).

4.2.2.2.3 Community Structure

Since metabarcoding can generate data on whole communities more efficiently than other methods, the cost and time for sample collection and identification do not limit the scope of taxa that can be included in an assessment like they do for morpho-taxonomic assessments (Pawlowski et al., 2018). As such, indices do not need to be reduced to a single value, but whole communities can be used to assess ecological status (Ruppert et al., 2019; Stoeck, Pan, et al., 2018). Such approaches are not commonly used in regulatory frameworks, but have potential to account for additional biotic factors, such as co-occurrence (Cordier, 2020). Community analyses that can potentially be used to assess ecological status are discussed further in the *Community Analyses* section.

4.2.2.3 Recommendations

When reporting on bioindicator or biotic index analyses, it should be noted what data were used for indicator analyses (presence/absence vs. quantitative), what markers were used, and whether data from different markers were combined, the reference database used and date of use, and any data transformations or normalization used to ensure consistency in methods where relevant. This information is required to interpret sources of bias in the results and enable cross-study comparisons using indices. Any newly developed indices using environmental genomics data need to be validated and should only be used to assess ecosystem status in environments and for stressors for which they have been validated.

4.2.3 Community Analyses

Metabarcoding generates data on a wide range of taxa, often across several major taxonomic groups. As such, the goal is often to compare whole communities across spatial or temporal gradients or across conditions, rather than focusing on individual species or taxonomic groups. While bioindicator taxa and biotic indices can provide a streamlined approach to assess ecosystem status, they rely on a limited scope of taxa and are designed to provide information on a specific measure of quality or status (Pawlowski et al., 2018). If work is being conducted where relevant indicators/indices are available for the environment and stressor of interest and the scope is limited to the stressor of interest, using indicators and/or an index may provide the most efficient approach to monitoring. However, where the

ecological information for an index is lacking, where multiple stressors are involved, where multiple taxonomic/functional/habitats are of interest and/or where ecosystem structure and function are of interest, whole community analyses are required. Below we have listed the most common approaches to community analysis, with their respective advantages and limitations for use with metabarcoding data and its unique properties (i.e., compositional, sparse, over dispersed; (Leite & Kuramae, 2020)). The analyses listed below can be conducted using taxonomic information (e.g., at the species-level) or they can be conducted using unique sequences or sequence clusters (i.e., MOTUs) as the taxonomic unit (Taberlet et al., 2018a). Using sequences as taxonomic units generally allows a larger portion of the data to be used but they are different units of measure than traditional species observations and thus results must be interpreted accordingly (Cordier et al., 2021). For example, MOTUs created based on a given similarity threshold do not necessarily correspond to a single species, genus, or taxonomic group. They may correspond to multiple species or conversely, two MOTUs may correspond to a single species. The understanding that MOTUs do not directly correspond to traditional taxonomic levels is essential to interpreting results based on MOTUs (Taberlet et al., 2018a). When metabarcoding data is analyzed using taxonomic assignments, these data can be biased due gaps in reference database information, leading to a large proportion of data going unused (Cordier et al., 2021). Note that some of community metrics and analyses described below are being integrated into complex indices (as noted in the *Biotic Indices* section).

4.2.3.1 Community Metrics

Community biodiversity metrics are values that provide a measure of biodiversity within, between, and across sampling locations. These values are not directly linked to any ecological status or environmental conditions but can be used to compare community biodiversity across conditions or gradients (Cordier et al., 2021). Various metrics exist to measure different aspects of biodiversity at different scales, including alpha diversity which measures diversity within a location, beta diversity which measures differences in biodiversity between locations, and gamma diversity which measures diversity across locations. These metrics are frequently used with metabarcoding data (e.g., (Brantschen et al., 2021; Doi et al., 2021; Leduc et al., 2019; Y. Li et al., 2018; Mauffrey et al., 2021)). Below we summarize the most commonly used metrics and their limitations, as well as recommendations for their use with metabarcoding data.

4.2.3.1.1 Alpha Diversity

Alpha diversity metrics measure the richness (number of different taxonomic units) and evenness (relative abundance of each taxonomic unit) of a community at a local scale, with metrics ranging from those that simply count the number of taxonomic units (richness) to those with an increasing reliance on the proportional abundances of taxonomic units (Daly et al., 2018). The most frequently used alpha diversity indices include, in order of increasing reliance on abundance information: observed richness, Shannon index, Simpson index, and Pielou's evenness (e.g., (Cordier et al., 2021; Foulon et al., 2016; Leduc et al., 2019; Zhou et al., 2022)). These can all be considered classical indices and are the most widely used largely because they are the oldest and simplest diversity indices available (Daly et al., 2018).

Richness is the simplest of these measures, requiring only a count of taxonomic units. Although this concept has been extended to include estimates of true richness, which include unseen species. For example, Chao1 (based on abundance) and Chao2 (based on incidence) are often used to estimate the number of undetected species and thus generate a estimate of true richness (Bukin et al., 2019; Gotelli

& Colwell, n.d.; C. Yang et al., 2014). Richness estimates are heavily impacted by rare species, making these metrics prone to bias when applied to metabarcoding data where sequencing and PCR errors can lead to spurious, low abundance, and low frequency sequences (Chiu & Chao, 2016). New methods are emerging that address this challenge and are introduced below.

Classical diversity metrics (e.g., Shannon index, Simpson index) are often calculated based on metabarcoding data, but these metrics can be complicated by uncertainties in taxonomic units as discussed above in this section and quantification discussed in the *Quantitative Analyses* section. However, it is generally accepted that these metrics can be used to calculate the statistical significance of a change in diversity following a disturbance and are frequently used in this context with metabarcoding data (e.g., (Mauffrey et al., 2021; Pawlowski et al., 2014; Pochon et al., 2015)).

Diversity metrics based on effective species numbers are a more recent approach to diversity measurement and they have the advantage of having more easily comparable and interpretable units (i.e., an effective number of species). The most commonly used framework for calculating effective species numbers is the Hill Numbers, which have the added advantage of being a parametric family of indices with a q parameter (known as the order) than can be adjusted to weight the relative contribution of different taxonomic units by their relative abundances (Daly et al., 2018). A q = 0 is species richness, q = 1 is analogous to the Shannon index and q = 2 is analogous to the Simpson index (Mächler et al., 2021). The values can be compared to one another to evaluate the contributions of taxonomic units based on relative abundance. The Hill Number framework is increasingly being used ~~to~~ with metabarcoding data due to the relative ease of interpretation and comparison across these different orders (e.g., (Doi et al., 2021; Mächler et al., 2021; Suter et al., 2021)).

The Hill Number framework is under active development and several extensions to this framework have been developed recently that provide additional functionality in interpreting metabarcoding data. Chiu and Chao 2016 developed a non-parametric estimator of the true singleton count to improve alpha diversity estimates from metabarcoding that have spurious singletons. Hill Numbers can also be calculated without rarefying data to control for sampling effort (e.g., sequencing depth) through coverage-based rarefaction and extrapolation (Chao & Jost, 2012; R. K. Colwell et al., 2012). This avoids the loss of useful data when rarefying to equal sampling effort. Furthermore, Hill Numbers can be calculated using phylogenetic or functional distances in the form of a species distance matrix (Alberdi et al., 2020; Taberlet et al., 2018a). Including phylogenetic information accounts for evolutionary relationships, reduces uncertainty regarding thresholds used to designate taxonomic units, and helps account for PCR/sequencing errors (Cordier et al., 2021; Tedersoo et al., 2022).

Classical indices continue to be widely used however, these new tools are increasingly being implemented (Roswell et al., 2021) and we expect their use will continue to become more popular as they overcome some challenges associated with the unique properties of metabarcoding data and are under active development. Alpha diversity metrics are scale dependent (Chase et al., 2019) and should only be compared across samples collected at the same scale and using the same methodology.

4.2.3.1.2 Beta Diversity

Beta diversity measures the variation in identities of taxonomic units among samples or sites (Tuomisto, 2010b). There are a wide range of beta diversity indices in use that use different data types (presence/absence/incidence data and/or abundance data) and focus on different aspects of beta

diversity (directional turnover or non-directional variation) (Andersen et al., 2011). Several articles have thoroughly reviewed the range of indices available (Andersen et al., 2011; Barwell et al., 2015; Koleff et al., 2003). Here, we focus on those most commonly used and relevant to metabarcoding studies, however there are many other options available that may be more suitable to a given monitoring or research objective. The most commonly used indices for metabarcoding include the pairwise dissimilarity measures: Jaccard index and Sørensen index (based on incidence data) and Bray-Curtis (based on abundance data) as well as the phylogeny based UniFrac (Galloway-Peña & Hanson, 2020; J. Liu & Zhang, 2021; Macher et al., 2018). Bray-Curtis can be calculated using either abundance or incidence data, and when calculated with incidence data it is the same as the Jaccard index (Galloway-Peña & Hanson, 2020).

The Jaccard and Sørensen indices are widely used similarity measures, however they have known limitations. These metrics ignore the relative magnitude of gains or losses in taxonomic units, or in other words, they are highly dependent on the overlap between two samples/sites but give less weight to unique detections in each community (Koleff et al., 2003). Additionally, these indices were shown not to accurately reflect differences in the dominant and rare across all scenarios, leading to biased estimates in some cases (Barwell et al., 2015). Despite these limitations, these two indices are useful for measuring turnover between communities and measuring variation among communities and they remain the most frequently used measures of dissimilarity for presence-absence data. Given the known biases associated with quantitative data generated using metabarcoding, they are widely applied in environmental genomics studies. Additionally, extensions to Jaccard and Sørensen indices are available to estimate beta diversity while accounting for unseen species by using abundance data or replicated incidence data (Chao et al., 2004).

Bray-Curtis, when applied to abundance data, gives a measure of dissimilarity that is most useful for comparisons of community composition and relative abundance (Andersen et al., 2011). Bray-Curtis performs well across a range of conceptual and sampling properties compared to other abundance based beta-diversity metrics (Barwell et al., 2015). While Bray-Curtis is more commonly used for metabarcoding data, the Morisita-Horn Index (Jost, 2006) performs equally well for abundance data (Barwell et al., 2015; Lim et al., 2016). Bray-Curtis dissimilarity requires even sampling, and thus may not identify community differences unless data is rarefied (Leite & Kuramae, 2020).

UniFrac is a widely used beta-diversity metric for bacterial microbiome studies using the 16S amplicon (Xia & Sun, 2017). This approach uses phylogenetic information, derived from 16S DNA sequences, to measure community similarity while accounting for phylogenetic relationships and can be calculated using presence/absence(unweighted) or abundance (weighted) data (Lozupone & Knight, 2005). Phylogenetic distance measures of beta diversity can provide more power for detecting community change by using the divergence between different sequences (Lozupone & Knight, 2005). While this is widely applied to microbiome studies, it does not apply well to other taxonomic groups since the marker genes used to study other taxonomic groups do not necessarily contain enough information to reveal phylogenetic relationships (Hajibabaei et al., 2007). As such, this approach is limited to cases where the same gene can be used for to resolve taxonomic units and evolutionary relationships. UniFrac values are known to be impacted by sampling effort, so appropriate methods to control or account for sampling effort should be applied when using UniFrac (Lozupone et al., 2011). Additionally, UniFrac values are less sensitive to changes in moderately abundant species and tend to be biased towards rare lineages or to

the most abundant lineages (Xia & Sun, 2017). Extensions to the UniFrac calculation have been developed to overcome these biases (Chang et al., 2011).

A new extension to the Hill Number framework was recently developed to calculate beta diversity using order number (q) to use incidence data ($q = 0$) or abundance data ($q > 0$) (Chao et al., 2023). As this is a new development it has yet to be widely applied and tested. However, given the advantages of the unified effective species number framework for alpha diversity metrics described above, extending this to include beta diversity may be very useful in accounting for the unique properties of metabarcoding data.

Beta diversity is often visualized and statistically tested using distance- or model-based approaches that are discussed further in the *Distance-based Analyses* and *Model-based Analyses* sections below.

4.2.3.1.3 Gamma Diversity

Gamma diversity, the total diversity at large-scale or across a landscape, can be summarized using the same metrics as alpha diversity (Tuomisto, 2010b) and is most often reported in metabarcoding literature as richness (e.g, (He et al., 2022, 2023; Keck et al., 2022)). However, it can also be calculated in an effective species number framework, such as Hill Numbers (Tuomisto, 2010a) thus generating values that are comparable to alpha diversity metrics generated in the same framework and more easily interpreted as effective species numbers.

4.2.3.1.4 Recommendations

Given known quantitative biases with metabarcoding data, there is debate about the use of quantitative or presence/absence data for diversity metric calculations. Research has shown that diversity indices calculated using abundance data give more robust results by decreasing the impact of rare taxonomic units (Taberlet et al., 2018a). Beta diversity estimates are less biased when using sequence abundance data (Barwell et al., 2015). Furthermore, when calculating beta diversity abundance data can be informative to estimate the effect of unseen taxonomic units (Chao et al., 2004). Since abundance data may generate more robust results, we recommend using an approach where diversity metrics are calculated with and without abundance data using a unified framework where these values can be directly compared when possible (e.g., Hill Numbers). This enables easy comparison of trends across samples and/or sites using metrics that include and do not include abundance data (e.g., (Suter et al., 2021)). Transformations are often recommended for quantitative analyses using metabarcoding (see *Quantitative Analyses* section). If quantitative data are used to calculate diversity metrics, results from different markers should be analyzed separately. Additionally, if analyses are being conducting at the level of sequence units (vs. taxonomy), different markers should be analyzed separately.

Diversity metrics are generally sensitive to sampling effort (Stier et al., 2016). Where variation in sampling effort exists, user should select an approach that accounts for sampling effort or is robust to variation in sampling effort (e.g., (Bennett & Gilbert, 2016; Cardoso et al., 2009; Chao & Jost, 2012; R. K. Colwell et al., 2012)) or apply other methods of controlling for sampling effort (e.g., rarefaction) (see *Controlling for Sampling Effort* section).

4.2.3.2 Distance-Based Analyses

Community data are often used to compare the abundance and distribution of organisms across environmental gradients or conditions, expanding on beta diversity comparisons and linking them with environmental variables. Community data have many properties that make analysis challenging, including intercorrelations between variables (i.e., biotic interactions between taxonomic units), non-normal probability distributions, and particularly relevant for metabarcoding studies, lots of zeros and more taxonomic units than sites (Jupke & Schäfer, 2020). Distance-based ordination uses a distance metric (or dissimilarity metric) to summarize multivariate data in a low dimensional form (e.g.., species by site matrices collapsed to dissimilarities between sites), then the distances/dissimilarities between communities can be related back to environmental variables (Roberts, 2020). There are a variety of metrics available, each with their properties and assumptions. A distance metric and ordination method must be chosen to best suit the data and research or monitoring goals (Jupke & Schäfer, 2020). However, often data do not meet the assumptions of a chosen metric or method and data are coerced into a format that better meets these assumptions (e.g., transformation, rarefaction) (Leite & Kuramae, 2020). Distance-based analyses have been the primary method for analysing community data since they were introduced (Roberts, 2020), and they are the most frequently used method for analyzing metabarcoding data (Leite & Kuramae, 2020).

There are a wide variety of distance metrics and ordination approaches available. We highlight those that are used frequently with metabarcoding data. The most common distance metrics used for metabarcoding data are Jaccard and Bray-Curtis (discussed in *Beta Diversity* section above). The most common ordination methods are non-parametric multidimensional scaling (NMDS) and principal coordinates analysis (PCoA), which are typically paired with permutational analysis of variance (PERMANOVA), analysis of similarities (ANOSIM), or Mantel test for statistical comparisons (Fujii et al., 2019; G. Jeunen et al., 2019; Krah & March-Salas, 2022; C. V. Robinson et al., 2022; Staehr et al., 2022). These analyses may be paired with post-hoc tests to determine the contribution of individual taxonomic units (e.g., SIMPER (Suter et al., 2021)). Another common exploratory approach is cluster analysis, which is used to separate communities into groups based on their similarity scores (e.g., UPGMA (Garcia-Vazquez et al., 2021; G. Jeunen et al., 2019); Ward's method (Stefanni et al., 2018)). Indicator species can be identified for the different cluster or groups (e.g., (Hajibabaei et al., 2019; G. Jeunen et al., 2019; K. M. West et al., 2020)). Methods of visualizing environmental variables together with community data in an ordination are also often used (e.g., envfit (Tapolczai et al., 2021)). Constrained ordinations aimed at directly assessing specific environmental gradients of interest are also used, mainly canonical correspondence analysis (CCA) and redundancy analysis (RDA) (e.g., (Cobo-Díaz et al., 2019; Huo et al., 2020)).

General considerations and limitations for using distance-based ordination methods are presented elsewhere (Austin, 2013; Paliy & Shankar, 2016; Ter Braak & Šmilauer, 2015) and there is no single method that will apply in all scenarios. Distance-based methods are still under active development and new tools continue to emerge that overcome limitations of pre-existing methods (e.g., t-SNE (Roberts, 2020)). The appropriate approach must be selected given the goals and properties of the data generated. There are several properties that apply to all metabarcoding datasets. First, the data are compositional in nature, meaning observation of each taxonomic unit within a sample are not independent of one another, which is an assumption of many statistical tests (Paliy & Shankar, 2016; Tedersoo et al., 2022). A range of data transformation options are used to address this although there is

no consensus around on a single transformation approach (e.g., (Banchi, Pallavicini, et al., 2020; Frankenfeld et al., 2022; C. V. Robinson et al., 2022; Tapolczai et al., 2021)). Second, metabarcoding data contains a lot of zeros, since many taxonomic units are only detected at a single site, further exacerbated by sequencing artefacts (Gold, Shelton, et al., 2023). Setting minimum read thresholds or minimum sample thresholds to remove taxonomic units that not detected across more than one sample or by a given number of sequences are common strategies to reduce the number of zeros in the data (e.g., (Shirazi et al., 2021)). There is no consensus in how these thresholds are applied for metabarcoding data and thresholds are often selected based on the properties of the data set being analyzed.

There are several limitations to distance-based community analyses that have been raised. First, they reduce multi dimensional data into a distance matrix and do not retain any information on individual taxonomic unites present across samples (Roberts, 2020). Second, they do not account species specific mean-variance relationships (Warton et al., n.d.). Finally, distance-based metrics make a lot of assumptions and therefore, data are often made to fit the assumptions of the analytical approach instead of using models to understand the variation and unique properties of the data (Leite & Kuramae, 2020). Some of these limitations are being addressed through model-based approaches described below, however comparisons of model-based and distance-based approaches show that distance-based approaches are robust compared to the current model-based approaches available (Jupke & Schäfer, 2020; Roberts, 2020). As model-based tools for community analyses continue to expand updated comparisons will be needed.

4.2.3.2.1 Recommendations

An appropriate distance metric and ordination approach must be selected based on the properties of the data being analyzed as well as the research and monitoring goals. Data should meet the assumptions of the chosen approach. If data transformations are used to meet these assumptions, there may be biases introduced through those transformations (Leite & Kuramae, 2020), thus selecting a metric and data transformation approach is a trade-off that will depend on the research and monitoring goals. If quantitative data are used to calculate diversity metrics, results from different markers should be analyzed separately. Additionally, if analyses are being conducting at the level of sequence units (vs. taxonomy), different markers should be analyzed separately.

Distance-based analyses are based on dissimilarity metrics, which are generally sensitive to sampling effort (Stier, Bolker, and Osenberg 2016). Where variation in sampling effort exists, users should apply a method of controlling for sampling effort (e.g., rarefaction) (see Controlling for Sampling Effort section), use a method that is robust to sampling effort (Beck et al., 2013) or demonstrate that communities have been sampling to an equal coverage despite variation in effort (i.e., all communities have been sampled to 95% coverage), if applicable (Chao & Jost, 2012).

4.2.3.3 Model-based Analyses

Model-based approaches to community analysis provide a means to jointly model multiple response variables (i.e., taxonomic units) and multiple fixed and random predictor variables (i.e., environmental conditions, study design) without reducing the dimensionality of the data (Warton, Blanchet, et al., 2015). Model-based approaches for community modeling are rapidly evolving with many new tools becoming available over the last 10 years. We highlight two modeling frameworks that have emerged

and are being used with metabarcoding data and discuss the advantages and limitations of these approaches.

4.2.3.3.1 Joint Species Distribution Models (JSDM)

Joint species distribution models make inferences at the community level by jointly modeling individual taxa distributions while acknowledging that taxa respond jointly to environmental conditions (Tikhonov et al., 2020). Modeling multiple taxa together improves predictions compared to single taxon models by using structure across taxa (Warton, Blanchet, et al., 2015). Joint species distribution models have been used to make inferences based on metabarcoding data and they can accommodate taxonomic data and well as molecular operational taxonomic units (Abrego et al., 2020; Fukasawa et al., 2022; Kačergytė et al., 2023; Tikhonov et al., 2020), however these models are yet to see widespread use in the environmental genomics community.

Ordinations can be conducted using a modelling approach with latent variables, such as a joint species distribution model. This model-based approach to ordination allows the compositionality of metabarcoding data to be accounted for using a site effect and the axes of the ordination use a probability distribution specified by the user to capture the variability in the dataset being used (Leite & Kuramae, 2020; Warton, Foster, et al., 2015). This approach can also estimate correlations across taxa, which is not possible with distance-based approaches that do not retain individual information on individual taxonomic units (Roberts, 2020; Warton, Blanchet, et al., 2015). There are many different frameworks available for running latent variables models, but those that have been used for metabarcoding data generally fall within joint species distribution models and provide additional relevant functionality (e.g., HMSC (Tikhonov et al., 2020)).

Additional information can be included in joint species distribution modeling to improve performance and inference, including phylogenetic and trait data (Ovaskainen & Abrego, 2020; Tikhonov et al., 2020). These data cannot necessarily be captured in distance-based approaches. Since this modeling approach is new, there are relatively few examples where joint species distribution models have used to their full capacity, including trait and phylogenetic information (but see (Abrego et al., 2022)). However, there is potential to see an increase in uptake of this approach since it accounts for many properties unique to community metabarcoding data.

4.2.3.3.2 Models for Imperfect Detection

While JSDM can account for a lot of environmental and design factors, they do not account for imperfect detection. Imperfect detection can occur at multiple levels in metabarcoding data (i.e., at the level of biological replicates and at the level of technical replicates) (Ficetola et al., 2015). Hierarchical species occupancy models were developed to analyze species distributions when the probability of detection or capture is less than 1 (Willoughby et al., 2016). By accounting for false negatives at multiple levels as well as environmental variables, these models generate robust ecological conclusions and can be used to inform optimal sampling design and methodology by improving our understanding of eDNA capture and detection probabilities (McClenaghan, Compson, et al., 2020). These models have been widely applied in single-species studies using eDNA and have started to be used for multi-species metabarcoding studies (Doi et al., 2019; McClenaghan, Fahner, et al., 2020). Calculating detection probabilities also generates a metric by which the reliability of the results can be assessed (Ficetola et al., 2015). A thorough explanation of occupancy models is included as Appendix A.

While accounting for imperfect detection can greatly improve estimates of species occurrence, there are limitations to this framework. Occupancy modelling is less robust when the probability of detection is relatively low and the number of replicates is also low (Willoughby et al., 2016). Sufficient replication is needed for accurate estimates with low detection probabilities, and low detections probabilities are not uncommon when using metabarcoding to assess whole communities (Hestetun et al., 2021). There is potential for false positives to be incorporated into occupancy models, but thus far only in single-species models and even so the data required to conduct an analysis with false positives may be challenging to obtain (e.g., deploying a secondary survey method) (Lahoz-Monfort et al., 2016). Species occupancy models do not account for correlations between species like JSDM, however models have recently been developed that incorporated JSDM and imperfect detection (Tobler et al., 2019). This integrated approach is likely to become more common but has not yet become widely used.

Hierarchical models that use abundance (N-mixture models) rather than presence-absence data can be fit using the same framework. Quantitative metabarcoding data has started to be used in this framework (Gold, Kelly, et al., 2023), although this framework is most often used with presence-absence data. Using quantitative metabarcoding data in hierarchical modeling is subject to the biases discussed in the *Quantitative Analyses* section and require careful interpretation since read counts are not the same units as individuals.

4.2.3.3.3 Recommendations

Model based approaches have the advantage of directly modeling the data and its properties and generating results for individual species and whole communities. However, models are more complex to fit, needing to estimate multiple parameters for taxonomic units and samples, and computationally intensive (Roberts, 2020). Generally, computational resources won't be a limiting factor for analyzing metabarcoding data as significant computational resources are already required for other steps in the bioinformatic workflow. However, for whole community comparisons (i.e., community-level change) models have not outperformed distance-based methods in simulation tests (Jupke & Schäfer, 2020; Roberts, 2020). As such, consideration of study goals will need to guide the choice of approach used for community analyses. Where high confidence detections units are needed, hierarchical occupancy models will provide a measure of reliability and can also inform robust sampling designs. Where community-level change that accounts for species interactions, traits, and/or phylogeny is desired, a joint species distribution model is recommended. Both modeling frameworks introduced here are under active development and model-based tools will likely increase in speed, accuracy, and ease-of-use. As the field evolves continued testing and comparison of approaches to community analyses will be required. These two modeling frameworks enable the use of common metrics to assess the power of these models to detect change in communities. For example, HMSC provides explanatory and predictive power metrics to assess the performance of the model across individual species or averaged across the whole community (Ovaskainen & Abrego, 2020). Power and trends are measured on a species-by-species basis in both modeling frameworks which distinguishes them from distance-based approaches which generate a single metric for the whole community. Given this distinction estimates of power cannot be compared between model-based approaches and distance-based approaches, however simulated data can be used to compare methods and determine each approach's sensitivity to community change (e.g., (Roberts, 2020)).

4.2.3.4 Networks & Food Webs

Network analysis is an increasingly popular tool for community analysis, including metabarcoding-based analyses, and can range from (relatively) simple co-occurrence networks to complex food webs that incorporate trophic and trait data (e.g., (Compson et al., 2018; Horn et al., 2019; Lanzén, Dahlgren, et al., 2021; C. V. Robinson et al., 2022)). Network analysis pairs well with metabarcoding data because metabarcoding can generate data on a broad range of organisms in an ecosystem (Compson et al., 2020). Networks can be constructed based on presence or abundance data, though read abundance data are generally scaled by sequencing depth (see *Quantitative Analysis* and *Controlling for Sampling Effort* sections) (Ritter et al., 2021).

Networks, even basic co-occurrence networks, can generate a wide variety of metrics to assess community structure (e.g., modularity, nestedness, diameter, average path length, transitivity, connectivity, etc.) (D'Alessandro & Mariani, 2021; Fais et al., 2020; Ritter et al., 2021; Tedersoo et al., 2022). These metrics provide a summary of a complex community and have been shown to reflect anthropogenic impacts on ecosystems using eDNA-based networks (Lanzén, Dahlgren, et al., 2021). This has led to suggestions that network metrics could be used as indicators of ecosystem function and/or integrity for biomonitoring and could be used as global indicators for ecosystem status (Compson et al., 2020; Cordier et al., 2021). Despite the promise of network analysis to provide broad-scale metrics of ecosystem status, we do not currently have enough information or understanding of the interactions between network properties, ecosystem function, and response to stressors, disturbance, or impact (Barroso-Bergadà et al., 2021; Clare et al., 2019; Cordier et al., 2021; C. V. Robinson et al., 2022). The ecological implications of network properties from eDNA remain difficult to interpret (Lanzén, Dahlgren, et al., 2021) and networks are sensitive to the compositionality of data, which applies to metabarcoding data, further complicating interpretation (Tedersoo et al., 2022). Additionally, molecular taxonomic units must be reconciled across studies before metrics based on unique sequences can be applied at a global scale (D. M. Evans et al., 2016). Further research is needed to understand how stressors and disturbances affect network properties and how changes in eDNA-based network properties should be interpreted before networks can be applied as indicators of ecosystem status or impacts.

Much of the research on eDNA-based networks has focused on co-occurrence networks, however, depending on whether taxonomic identifications or molecular taxonomic units are used to create these networks, additional layers of data can be integrated into networks. For example, trophic information can be integrated to create food webs if networks are created with taxonomic identification (D'Alessandro & Mariani, 2021). Adding trait data can be very time intensive, though new tools are being trialled to facilitate this process (e.g., machine learning (Compson et al., 2018)). For microbial networks created using sequence data, both functional data and phylogenetic data have been included in networks (D. M. Evans et al., 2016; Z. Liu et al., 2021). While incorporating more data should generate more robust and reliable networks, more research is needed to inform the interpretation of complex networks generated with metabarcoding data.

4.2.3.5 Reporting Recommendations for All Community Analyses

To enable appropriate interpretation, reporting for all community analyses should note whether quantitative or incidence-based data was used to calculated metrics, what markers were used, if data from different markers were combined, any data transformations used, and the taxonomic unit used for metric calculations (e.g., species or molecular operational taxonomic unit). Metabarcoding data is often

collected using replicated samples (biological or technical replicates) and community analyses can be calculated with replicates combined or separate. Replicates are not independent observations so interpretation must consider the unit of measure used for the calculation of community metrics. Reporting should note how data from biological and technical replicates was handled. Across many community analyses, low prevalence taxonomic units are removed to reduce noise and increase the power to detect community level patterns. If applied, the threshold used to remove lore prevalence taxa should be reported.

4.3 Field Metadata

Collecting the appropriate metadata during sample collection is essential for the effective interpretation of eDNA data. In addition to facilitating interpretation, metadata standards allow for better data integration and interoperability, reproducibility, quality, and collaboration (Field et al., 2008; Kimble et al., 2022; Yilmaz et al., 2011). Alongside the use of metadata standards, the use of metadata management software can lead to increased structure and help prevent idiosyncratic entries by reducing the number of manual entries that can lead to incompatible data and decreased data integrity (Kimble et al., 2022).

There are multiple standards available describing essential and recommended metadata fields that should be recorded when analysing eDNA depending on the type of analysis is performed. The most relevant and widely used is the MIxS (Minimum information about any (x) sequence) framework (Yilmaz et al., 2011) compiled by the Genomic Standards Consortium, which provides guidelines for the minimum information about any type of sequence. This framework provides guidelines for the minimum information required for any type of DNA sequence and includes more specific guidelines for certain types of analysis, including marker gene sequencing (MIMARKS), metagenomes (MIMS), and genome sequencing (MIGS).

The shared descriptors in these frameworks include sample descriptors, temporal and geospatial descriptors, and technical metadata listed below:

- Project Name
- Sample Name
- Sample Size
- Collection Date
- Geographic Location (Country/Sea/Region)
- Geographic Location (latitude and longitude)
- Broad-scale Environmental Context (i.e., biome)Local Environmental Context (e.g., cliff, harbour, etc.)Environmental Medium (e.g., soil, sediment, seawater)
- Sequencing Method

Required and recommended metadata fields specific to each analysis workflow are also provided. There is also a framework for metadata related to Quantitative Real-Time PCR (qPCR) experiments (MIQE) (Bustin et al., 2009) that provides standards for metadata when qPCR techniques are used to analyse eDNA samples.

Additional metadata about the sampling environment should be also recorded to provide a robust basis for analysis, interpretation, and comparability. There are several environmental packages defined by the

MIxS framework that include water, soil, sediment, and air among others, which are the most relevant to conventional eDNA sampling. These packages include fields and descriptions for commonly collected metadata pertinent to the environment from which samples were collected (e.g., chlorophyll, conductivity, or dissolved oxygen from a water sample).

5 Metagenomics

Amplicon sequencing has revolutionized biodiversity assessment by enabling the identification of novel organisms based on their DNA sequences. However, the lack of culture representatives for many groups (i.e., those that could be easily cultured within the laboratory), such as Archaea, demanded a new approach. In 1996, Stein et al., reported the first attempt to address this problem through random shotgun sequencing of archaeal clones extracted from picoplankton assemblage collected in the Pacific Ocean. Two years later, the term "metagenome" was coined to describe "the collective genomes of soil microflora" (Handelsman et al., 1998). Since then, "metagenomics" have been used to describe various data structures. Although the terminology surrounding metagenomics can be confusing, untargeted shotgun metagenomics provides a powerful tool for investigating the functional potential and taxonomic composition of environmental DNA.

In recent years, the reduced cost and improvement in DNA sequencing have enabled large-scale metagenomics to study global biodiversity (Sunagawa et al., 2020). This approach involves extracting total DNA from a sample, such as water, soil, fecal, biopsy, or swab, and preparing a sequencing library depending on the sequencing technology platform. Illumina (HiSeq, NextSeq, and NovaSeq) is currently the most common sequencing platform for metagenomic sequencing, generating 150-250bp sequence reads. PacBIO and Oxford Nanopore can sequence longer DNA fragments but are less frequently used due to the higher cost. The higher taxonomic and functional resolution of metagenomic sequencing has significantly improved our understanding of the biodiversity and provided insights into functional role of organisms in an ecosystem.

5.1 Metagenomics Quality Control and Filtering

When it comes to filtering primers and removing low-quality reads from raw metagenomics data, several crucial factors must be considered to ensure a robust approach. These factors encompass sequence quality, adapter removal, read length distribution, throughput speed, and the impact on downstream analysis. Selecting the most suitable tool depends on the quality of the input data and time constraints. Each tool typically employs slightly different default cut-offs, but these parameters can be adjusted manually within the respective tools. It is worth noting that the majority of existing tools are designed for the Illumina sequencing platform, which is currently the standard for metagenomic sequencing. By carefully considering these factors and the specific requirements of the project, one can make informed decisions about the appropriate tool to employ for their metagenomics data analysis.

Trimmomatic (Bolger et al., 2014) is a widely used tool for trimming Illumina sequencing reads. It provides various trimming options, such as removing adapter sequences, trimming low-quality bases, and removing reads below a certain length threshold. Cutadapt (C. Martin, 2011) is another popular tool for adapter trimming in metagenomics. It has a flexible algorithm that can handle a wide range of adapter sequences and provides various quality trimming options. BBDuk (part of the BBMap suite; Bushnell, 2014) is a comprehensive tool for quality trimming and adapter removal. It offers advanced options for handling complex adapter sequences and has built-in error correction capabilities. Fastp (S. Chen et al., 2018) is a relatively new tool that has gained popularity in the metagenomics community. It performs both adapter trimming and quality filtering and is known for its fast-processing speed.

With the ever-increasing sequencing throughput of platforms like Novaseq, the need for faster performance has become a critical consideration in metagenomic data analysis. We need to find tools that can maintain the same level of quality while significantly reducing processing time. In a recent benchmarking study conducted by Chen et al., 2011, various metagenomic trimmers, including FASTQC, Cutadapt, SOAPnuke, AfterQC, and Trimmomatic, were compared. The results of the study revealed that Fastp outperforms other tools in terms of performance speed. Additionally, when compared to other tools for adapter trimming, base correction, sliding window quality pruning, polyG and polyX tail trimming, Fastp consistently achieved either the same or improved quality outcomes (S. Chen et al., 2018). These findings highlight Fastp as an ultra-fast fastq processor and a robust tool for metagenomic quality control and filtering. Fastp is an open-source tool and is publicly available on GitHub, further enhancing its accessibility to researchers in the field.

5.2 Read-based metagenomics

Read-based metagenomics is a powerful tool for profiling the taxonomy and functional capacity of eDNA. This approach allows us to gain insights into the genetic material present in a sample, even if we do not know which specific organisms are contributing to it or if the genes are not annotated in publicly available databases. To achieve this, read-based metagenomics relies on comparing high-quality reads to external sequence databases using supervised learning methods. There are four main approaches for taxonomic assignment:

1. Similarity search: this method uses homology or alignment-based methods based on the lowest common ancestor (LCA) to compare the query sequence to databases. Examples of tools that use this approach include BLAST (Altschul et al., 1997) and MEGAN (Huson et al., 2011).
2. Composition methods: this approach uses k-mer counts or frequencies to compare the query sequence to databases. Examples of tools that use this approach include KRAKEN (D. E. Wood & Salzberg, 2014) and CLARK (Ounit et al., 2015).
3. Phylogenetic approach: this method uses evolutionary models coupled with homology-based or interpolated Markov models to compare the query sequence to databases. An example of a tool that uses this approach is Phymm (Brady & Salzberg, 2009).
4. Short-read mapping: In the short-read mapping strategy, sequences that have successfully met quality standards are aligned to a known reference genome. This approach is frequently employed in metagenomics to address targeted inquiries concerning the presence of genes or genomes, rather than aiming to comprehensively characterize community makeup. Among the software options available, two prominent contenders that have consistently outperformed alternative approaches are bwa (H. Li & Durbin, 2009) and bowtie2 (Langmead & Salzberg, 2012).

Each of these approaches has its strengths and weaknesses, and the choice of method will depend on the specific research question and available resources. The homology-based method, such as BLAST, is a commonly used approach that searching each query sequences against large databases. While this method is reliable in providing a robust taxonomic resolution, it can be computationally intensive and time-consuming, especially for deep shotgun metagenomics with millions of reads. The Megablast (Y. Chen et al., 2015) algorithm was developed as a solution to this issue, but for extremely large datasets, alternative strategies may be necessary to expedite the process.

The phylogenetic approach is an evolutionary-based method that uses maximum likelihood, neighbor-joining, or Bayesian methods to determine the appropriate placement of a query sequence on a phylogenetic tree (Bazinet & Cummings, 2012). This approach uses simple observation to find where an inserted branch is divergent from a node representing a species or higher rank. However, the phylogenetic-based methods can be computationally demanding as it involves multiple alignments, fixed topology (e.g., NCBI taxonomy), and the insertion of a query sequence into the reference alignment. These steps require significant computational power, and therefore, the phylogenetic approach may not be suitable for large-scale datasets.

Compositional methods, such as Naive Bayesian classifiers, interpolated Markov models (IMMs), and k-mer/k-nearest-neighbor algorithms such as Kraken2, and CLARK (Ames et al., 2013), offer faster computational speeds compared to alignment or phylogenetic-based approaches. However, it is important to note that these methods do require a substantial amount of computational memory since a pre-computed database needs to be loaded into memory beforehand.

Marker-based algorithms present another read-based analysis, employing a selected set of representative genes, or markers, instead of relying on an extensive database encompassing all known sequences for microbial composition profiling. These methods, which do not require genome assembly, have been successfully employed in taxonomic and functional analysis of large human-associated metagenomic datasets from the MetaHIT and HMP consortia. Notably, mOTU (Ruscheweyh et al., 2021) and MetaPhlAn (Manghi et al., 2023) have been utilized for this purpose. For instance, the application of clade-specific markers from the CHOCOPhlAn database, incorporated in MetaPhlan, has demonstrated accurate estimation of microbial composition, along with improved computational efficiency. However, it is important to note that profiling unknown microbes, particularly regarding gene families and functions, can be challenging. The HUMAnN package, commonly used for pathway and gene family profiling, often encounters 40% unmapped reads, as reported by Franzosa et al. (2018). While reference genome databases and marker accuracy continue to expand, issues such as incomplete or insufficient annotation of these databases persist.

5.2.1 Taxonomic classification

The Critical Assessment of Metagenome Interpretation (CAMI; Meyer et al., 2022; Sczyrba et al., 2017) challenge is a pivotal initiative that has significantly propelled the field of metagenomic analysis forward. By establishing a comprehensive benchmarking framework, CAMI offers a standardized platform to assess and compare the performance of diverse tools employed in the interpretation of metagenomic data. In this challenge, the same benchmarking datasets including marine and host associated metagenomic communities were used by participants to analyse the raw data using their respective metagenomic tool. Leveraging the results obtained from CAMI, we can impartially examine and scrutinize the capabilities and limitations of various metagenomic tools. This unbiased approach enables us to gain a deeper understanding of their effectiveness, and to make informed decisions when selecting the most suitable tools for the metagenomic analysis.

For users who have access to high-memory computational resources, Kraken classifiers provide reliable taxonomic estimations across a wide range of metagenomic datasets, including environmental DNA and host-associated microbiota. The Kraken package utilizes a k-mer based algorithm and is complemented by derivative tools such as Bracken, KrakenUniq, and Kraken2 (D. E. Wood et al., 2019; D. E. Wood &

Salzberg, 2014). One notable advantage of Kraken is its flexibility in creating custom databases, and it sets itself apart by delivering exceptional runtime performance compared to alternative tools.

In situations where high-memory computers are not available, Metaphlan (Manghi et al., 2023) which is a marker-based approach emerges as a recommended choice due to its impressive accuracy and minimal memory requirements. It should be noted, however, that Metaphlan does not support custom databases and is primarily tailored for analyzing microbial communities. As an alternative approach, Centrifuge (Kim et al., 2016) offers the ability to utilize custom databases when high-memory machines are inaccessible (Ye et al., 2019). Another highly accurate option is Megablast (Z. Zhang et al., 2000), which can leverage extensive databases such as the National Center for Biotechnology information (NCBI) non-redundant nucleotide (nt) database. However, a drawback of using Megablast is the extensive processing time, which can take weeks to complete the analysis of a single dataset. Currently, there exists a considerable gap in the completeness of reference databases for environmental DNA, particularly for metagenomic data requiring whole genomic references to cover the metagenomic reads. To tackle this challenge, we recommend utilizing large databases such as nt, which can be coupled with time-efficient tools like Kraken to ensure comprehensive analysis (Singer et al., 2020).

5.2.2 Functional annotation

BLASTP (Altschul et al., 1997) has long been considered the gold standard for protein alignment in metagenomics. However, similar to the challenges faced by BLASTN (Camacho et al., 2009) in taxonomic assignment, BLASTP suffers from time inefficiency in protein alignment. For instance, Buchfink et al., (2021) estimated that aligning the non-redundant protein (nr) database from NCBI against the UniRef50 database using BLASTP would require over two months on a cluster with 20,800 cores.

Several alternatives to BLASTP exist, including MEGAN (Huson et al., 2011), DIAMOND (Buchfink et al., 2021), and MMSeq2 (Steinegger & Söding, 2017). However, DIAMOND stands out for its exceptional speed and accuracy. It delivers impressive performance and enables the use of custom databases, making it suitable for diverse datasets.

On the other hand, tools like MG-Rast (Meyer et al., 2008) and Humann (Franzosa et al., 2018) focus more on functional annotations using previously identified markers and provide annotations at the species level. For robust functional annotations across different functional levels and systems, Humann3 is a recommended tool. However, it is primarily designed for the analysis of microbial communities.

As an alternative to Humann, a viable approach for diverse datasets involves using DIAMOND for protein alignment, clustering the identified proteins, and utilizing a protein annotation tool such as EggNOG (Huerta-Cepas et al., 2017). It is important to note that read-based approaches for functional profiling, such as Humann, in metagenomics do not offer high functional resolution. Metagenomic reads often lack complete genes or functional pathways, leading to false-positive results. Therefore, the recommended approach for functional annotation in metagenomics is to employ assembly-based approaches, which will be discussed in the subsequent section.

5.2.3 Functional inferences using amplicon sequencing.

Metagenomic analysis provides direct functional insights but can be expensive, time-consuming, and requires greater computational resources. As an alternative, tools like Picrust (Langille et al., 2013) offers a different approach to functional prediction, circumventing some of the challenges posed by metagenomics. This approach utilizes amplicon markers, notably the 16S rRNA gene, as a basis for estimating the potential functions of microorganisms. This method depends on comprehensive databases of bacterial genomes that link 16S markers to an array of possible functions in different species. Following Picrust, the development of similar tools like Tax4Fun2 (Wemheuer et al., 2020), Piphillin (Iwai et al., 2016), and Picrust2 (Douglas et al., 2020) has further expanded the capabilities of this approach. These tools are particularly valuable for generating hypotheses in microbial studies in a cost-effective manner. By predicting the functional capabilities of microorganisms from a less detailed dataset, they provide an initial insight into the roles these microorganisms may play in their respective environments.

Despite the advantages of using Picrust as an alternative to direct metagenomic analysis, there are inherent challenges. The resolution of Amplicon Sequence Variants (ASVs) obtained through this method often does not extend to the species level. When it does, the functional potential can vary significantly among different strains within the same microbial group, posing a challenge in accurately characterizing microbial functions. Although broad functional categories like KEGG and COGG pathways might appear similar in both amplicon-based predictions and direct metagenomic analyses, the specifics of functional mapping can vary greatly. This is primarily due to the lack of strain-level resolution in amplicon-based functions, underscoring the need for subsequent detailed metagenomic studies for comprehensive functional understanding.

5.3 Assembly-based Metagenomics

In recent years, the cost of sequencing has significantly decreased, leading to a remarkable expansion in genomic data across various organisms. However, despite these advancements, existing sequencing technologies are limited to generating relatively small genomic fragments, typically ranging from 150 base pairs (e.g., Illumina) to approximately >10-20 kilobases (e.g., PacBIO). Considering that the size of a typical bacterial genome is around 5 million base pairs, the process of reconstructing the complete genome necessitates the utilization of sophisticated computational algorithms capable of assembling these short sequencing reads into a coherent whole.

In genome assembly, two primary approaches are utilized: reference-based and *de novo* assembly (also known as reference-independent assembly). Given the incomplete nature of genomic reference databases, it is crucial to employ an unbiased reference-free approach when reconstructing the metagenome structure of environmental and host-associated genomes. While there have been some efforts to utilize reference-guided methods (Dutilh et al., 2009; Lischer & Shimizu, 2017), *de novo* assemblers have predominantly been employed for the assembly of microbial genomes and metagenomics (Quince et al., 2017).

The *de novo* assembly approach in genome assembly can be categorized into three main categories: OLC graph, string graph, and de Bruijn graph. OLC algorithms, including Celera (Myers et al., 2000), AMOS (Treangen et al., 2011), and PCAP (Huang et al., 2003), operate by identifying overlaps among reads, constructing a layout graph based on these overlaps, and inferring consensus reads from the layout. String-based methods, such as SGA and FALCON (Chin et al., 2016), are derived from OLC graph-based

methods and aim to eliminate duplicate and substring reads before building graph layouts. De Bruijn graph is the most widely used *de novo* assembly framework, where reads are divided into k-mers representing nodes. Overlapping nodes with k-1 bases form arcs within reads, while k-mers sharing k-1 bases between reads create direct edges. De Bruijn graphs can be classified into Hamiltonian and Eulerian graphs. Hamiltonian graphs represent nodes as k-mers, and the edge represents the overlap (similar to OLC approach), while Eulerian graphs consider k-mers as edges. Eulerian-based algorithms like IDBA-UC (Peng et al., 2012) and SPAdes (Bankevich et al., 2012) are more effective for assembling large genomes compared to Hamiltonian-based algorithms like SOAPdenovo and Velvet (Zerbino & Birney, 2008), as they avoid a simplification step required in constructing the Hamiltonian path. The results of the CAMI challenge indicate that meta-SPAdes, a specialized variant of the SPAdes package designed for metagenomic data, outperforms other assemblers when dealing with large environmental datasets and human-associated microbiota. It has established itself as the current gold standard for metagenomic assembly.

5.3.1 Metagenomic binning: resolving genomes from metagenomics.

Metagenomic assembly generates thousands of contigs with varying lengths, but their origins and the number of genomes present in a community remain unclear. Unsupervised binning is a common approach for identifying metagenome assembled genomes (MAGs). Binning algorithms predominantly rely on tetranucleotide frequencies (Dick et al., 2009) and coverage information to identify similarities between contigs and cluster them together. Some widely used metagenomic binning algorithms include CONCOCT (Alneberg et al., 2014), MetaBAT (Kang et al., 2019), and MaxBin (Wu et al., 2016).

Genome-resolved metagenomics has facilitated the discovery of numerous microbial groups without representative cultures and significantly improved microbial genome collections (Nayfach, Roux, et al., 2021). However, assessing the quality of MAGs remains challenging. The current metrics, completeness, and contamination, based on single-copy core genes, lack sensitivity, and fail to evaluate the quality of the accessory genome (Parks et al., 2015).

5.3.2 Challenges and opportunities in *de novo* assembly metagenomics

De novo assembly of metagenomic data presents several challenges that stem from the complexities inherent in analyzing mixed microbial communities. These challenges include:

1. **Sequencing Errors:** Metagenomic datasets often contain errors introduced during sequencing, which can lead to incorrect base calls and complicate the assembly process. De Bruijn graph-based assemblers, like Meta-IDBA (Peng et al., 2011), MetaVelvet (Namiki et al., 2011), and metaSPAdes (Bankevich et al., 2012), are sensitive to these errors, potentially resulting in the generation of fragmented or erroneous contigs.
2. **Repetitive Regions:** Repetitive regions in genomes can cause ambiguity in the assembly graph. In metagenomics, this challenge is exacerbated as repetitive regions may arise from multiple distinct species or strains with similar genetic sequences. Resolving these regions accurately is difficult and can lead to fragmented assemblies.
3. **Uneven Genomic Coverage:** Unlike traditional genome assembly, metagenomic samples comprise a mixture of organisms, each with varying abundances. High abundance genomes are well represented in the data and are more likely to be accurately assembled. However, low

abundance genomes may have sparse coverage, making it challenging to reconstruct their full genomes, resulting in fragmented contigs.

4. **Highly Diverse Microbial Communities:** Metagenomic samples can contain highly diverse microbial communities, with varying levels of relatedness between species. As a result, assembling individual genomes from such complex mixtures becomes intricate and may lead to chimeric contigs or the merging of sequences from different organisms.
5. **Incomplete Genomes:** Due to the presence of rare or low-abundance species, some genomes may not be sufficiently covered to produce complete assemblies. This leads to gaps in the assembled contigs, hindering our ability to comprehensively study the microbial diversity present in the sample.
6. **Computational Resources:** The computational demands of *de novo* assembly for metagenomic datasets can be substantial, particularly for large-scale studies or when using long-read technologies. Assembling billions of reads and constructing complex graphs necessitates powerful computational infrastructure and sufficient memory.

To address these challenges in metagenomic *de novo* assembly, specialized methods and algorithms were developed that has a potential to improve and overcome the mentioned obstacles. These include:

1. **Abundance-Dependent Assembly:** Binning algorithms, such as MetaBAT (Kang et al., 2019) and MaxBin (Wu et al., 2016), group contigs based on their abundance patterns, aiding in the reconstruction of individual genomes from complex communities. By leveraging abundance information, these methods can improve the assembly of low-abundance genomes.
2. **Hybrid Assembly:** Integrating data from both short-read and long-read sequencing technologies (hybrid assembly) can help resolve complex regions and produce more contiguous assemblies. Long-read technologies, such as PacBio and Oxford Nanopore, are particularly valuable for spanning repetitive regions and improving contig continuity.
3. **Iterative Approaches:** Some assemblers, like IDBA-UD (Peng et al., 2012), use iterative strategies to refine the assembly graph, potentially leading to better contig lengths and quality, particularly for challenging metagenomic datasets.
4. **Reference-Guided Assembly:** Utilizing reference genomes, when available, can aid in the assembly process by anchoring reads to known sequences, potentially filling gaps in low coverage regions and providing a scaffold for the assembly.
5. **Error Correction:** Employing error correction tools, such as BayesHammer (Nikolenko et al., 2013), and BFC (H. Li, 2015), can help identify and rectify sequencing errors before assembly, improving the accuracy of the final contigs.
6. **Validation and Quality Assessment:** It is essential to validate the quality of assembled contigs using tools like CheckM (Parks et al., 2015) or QUAST (Gurevich et al., 2013), which assess the completeness and accuracy of the reconstructed genomes.

By combining these approaches and leveraging the strengths of various tools, we can mitigate the challenges associated with de novo assembly in metagenomics and gain deeper insights into the structure and function of complex microbial communities. However, given the diverse nature of metagenomic datasets, no single method may provide a complete solution, and selecting appropriate strategies based on the specific characteristics of the dataset remains crucial for successful metagenomic assembly.

5.3.3 A comprehensive genome resolved metagenomic pipeline for prokaryote and eukaryote organisms.

In this section, we present a robust metagenomic pipeline designed to gain insights into prokaryote, eukaryote, and viral genomes present within complex metagenomic datasets and implemented for genome-resolved metagenomic analysis in the Centre for Environmental Genomics Applications, located in St. John's, NL, Canada (see Figure 13).

5.3.3.1 Metagenomic assembly

We processed the raw shotgun reads by employing Fastp (S. Chen et al., 2018) to eliminate low-quality reads and Illumina adapters. Subsequently, we utilized metaSPADE (Bankevich et al., 2012) for the assembly of filtered reads, with contigs shorter than 1kb being removed to minimize misassemblies. To gain insights into the metagenome communities, we mapped the filtered reads to the remaining contigs using BWA-mem (H. Li & Durbin, 2009), generating a coverage information table. We adopted the assumption that reads originating from the same cell would have similar consistent coverage information and tetranucleotide frequencies, which led us to bin similar contigs together using Metabat2 (Kang et al., 2019). To assess the quality of these bins, we performed CheckM (Parks et al., 2015) and EukCC (Saary et al., 2020) analyses to identify single-copy core genes for prokaryote and eukaryote MAGs, respectively. Additionally, we calculated the total assembly length of each bin using a custom Python script. This comprehensive approach ensured the acquisition of accurate and reliable metagenomic contigs and bins.

5.3.3.2 Prokaryote MAGs

Bins meeting specific criteria were selected as prokaryote MAGs, requiring >70% completion, <10% contamination (CheckM), and a size of >2.5Mb to ensure they possessed the minimum required information. For taxonomic prediction of these MAGs, we utilized the GTDB-tk (Chaumeil et al., 2020) algorithm, which aligns GTDB markers against the genomes, facilitating the generation of alignments for phylogenetic trees. To further enhance our understanding of these MAGs, we employed Bakta (Schwengers et al., 2021) for CDS annotation and preliminary functional predictions. For comprehensive functional annotation, we ran eggNOG mapper2 (Huerta-Cepas et al., 2017) with DIAMOND and HMM models, enabling the prediction of Gene Ontology (GOs), Enzyme Commission number (EC), Kyoto Encyclopedia of genes and genomes (KEGG), and pfmas. This rigorous annotation process allowed us to gain insights into the functional characteristics and taxonomy of the selected prokaryote MAGs.

5.3.3.3 Eukaryote MAGs

We assessed the genomic composition of the remaining bins (not categorized as prokaryote MAGs) using EukRep (P. T. West et al., 2018). Bins meeting specific criteria were selected as Eukaryote MAGs, requiring >1Mb of Eukaryotic DNA (EukRep), >2.5Mb in length, >90% completion, and <10% contamination (EukCC) to ensure they contained sufficient and reliable information. To predict the taxonomy of each Eukaryote MAG, we employed KrakenUnique with the "nt" database from GenBank, enabling us to identify the species with the highest proportion of assigned k-mers. For gene prediction and protein annotation of the Eukaryote MAGs, we utilized the AUGUSTUS option from the BUSCO (Simão et al., 2015) packages. The predicted proteins were further annotated using the eggnog and MMseq2 databases, providing comprehensive functional insights into the selected Eukaryote MAGs. This

rigorous approach allowed us to gain a comprehensive understanding of the genomic and taxonomic characteristics of the Eukaryote MAGs in our pipeline.

5.3.3.4 Viral MAGs

The remaining contigs, which were not categorized as prokaryote or eukaryote MAGs, underwent further processing using the Phamb (Johansen et al., 2022) pipeline. This pipeline leverages DeepVirFinder (Ren et al., 2020) for the identification of viral contigs and VAMB (Nissen et al., 2021) for the binning of these viral contigs. The pipeline incorporates prodigal and a random forest model for accurate viral genome annotation. To ensure the quality and taxonomic classification of these viral bins, we employed CheckV (Nayfach, Camargo, et al., 2021) and DeepVirFinder. CheckV allowed us to evaluate the quality of the bins, while DeepVirFinder provided valuable insights into their viral taxonomy. By applying this approach, we aimed to gain a comprehensive understanding of the viral component within the metagenomic data and enrich our knowledge of the diverse viral communities present in the samples.

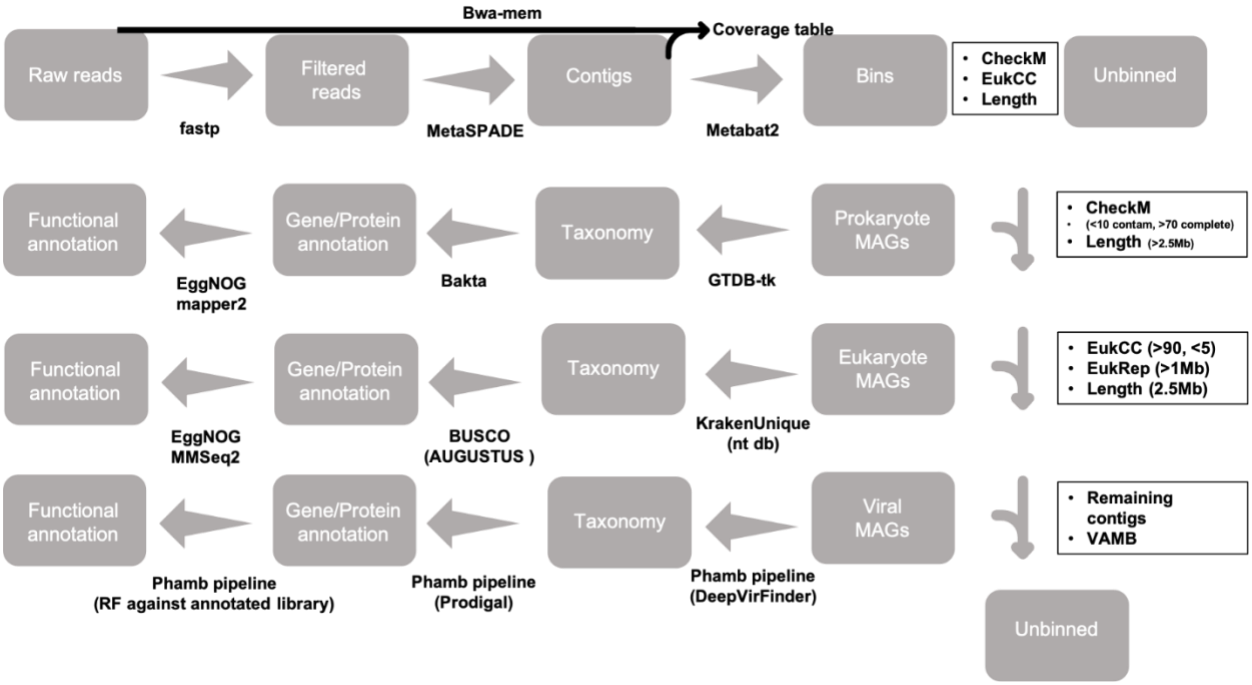


Figure 13: A comprehensive genome-resolved metagenomic pipeline

5.4 Metatranscriptomics

Metagenomics serves as a genetic census, addressing fundamental questions of community composition and functional potential. By scrutinizing the complete genomic content of a sample, metagenomics answers the queries "who populates the ecosystem" and "what activities are underway." This comprehensive survey captures the genetic blueprints of all organisms present, revealing their potential capabilities and providing a panoramic view of the genetic diversity within the community. However, it's essential to recognize that metagenomics also encompasses dormant and inoperative DNA segments, which can lead to an overestimation of functional potential.

In contrast, metatranscriptomics takes a dynamic snapshot of microbial activity, portraying the ongoing symphony of functional gene expression within a community. By focusing on the transcribed RNA molecules, metatranscriptomics delves into the question "what are they actively doing?" This approach captures the moment-to-moment activities of microbial players, highlighting which genes are actively involved in metabolic pathways, signal transduction, and other vital functions. Metatranscriptomics offers insights into real-time responses to environmental cues, revealing the true functional dynamics of the community.

The methodology underlying metatranscriptomics shares commonalities with metagenomics, primarily in the utilization of shotgun whole-genome sequencing. However, the key difference lies in the sequence material under investigation. In metatranscriptomics, RNA takes center stage as the starting point. The process involves extracting RNA, followed by the depletion of structural RNA to isolate the functional transcripts. These transcripts are then transformed into complementary DNA (cDNA), enabling the conversion of RNA-based information into a format that can be readily sequenced and analyzed. This process effectively captures the active genetic processes underway within the microbial community. Metatranscriptomics demands a more extensive sequencing effort compared to metagenomics when applied to the same microbial community. To put it in perspective, if the least common organism is around 100 times less prevalent than the most abundant one, then roughly 1% of the DNA reads associated with the less common organism should be substantial enough for its metagenomic detection.

Metatranscriptomics bioinformatics typically encompass a series of essential stages, comprising:

1. Quality control of shotgun sequencing data to eliminate or trim erroneous reads.
2. Alignment with reference sequences or de novo assembly for characterizing transcript abundance.
3. Functional and taxonomic characterization to discern active components and community members.
4. Application of statistical analyses to normalize expression and discern variations between distinct conditions, such as identifying differential expression or co-expression dynamics.

Earlier, we delved into steps 1-3 within the metagenomics, and it's worth noting that the same core methodology and concepts extend seamlessly to metatranscriptomics. There exist several refined workflows designed to seamlessly integrate these steps, as evidenced by prominent pipelines such as COMAN (Ni et al., 2016), SAMSA2 (Westreich et al., 2018), HUMAnN (Beghini et al., n.d.), and SqueezeMeta (Tamames & Puente-Sánchez, 2019). These workflows offer systematic and efficient frameworks to facilitate in-depth metatranscriptomic analyses. Remarkably, the HuMAnN workflow stands out as the most widely adopted pipeline for both environmental and host-associated metatranscriptomics studies. This pipeline's popularity underscores its effectiveness in enabling comprehensive insights into the intricate world of active genetic processes within microbial communities.

5.4.1 Statistical considerations and normalization strategies for metatranscriptomics

Much like single-organism RNA-seq, metatranscriptomics measurements rely on whole numbers representing sequencing reads. These counts are influenced by both the amount of sequencing performed on a sample and the length of the transcripts being studied. The counts often include numerous instances of zeros, which often signify instances where a transcript wasn't detected rather

than a complete lack of expression. Moreover, the levels of transcript abundance within a particular species can encompass a wide range of values, sometimes differing by several orders of magnitude. Similarly, a single transcript's abundance might vary substantially across different samples, leading to a deviation from the assumptions of normal distribution that underlie many common statistical tests. To tackle these challenges, various statistical techniques such as edgeR (M. D. Robinson et al., 2010), DESeq2 (Love et al., 2014), and NOISeq (Tarazona et al., 2015) have been developed for RNA-seq analysis.

Metatranscriptomics poses distinct normalization challenges compared to single-organism RNA-seq, chiefly tied to transcript-gene copy relationships. Unlike the typical assumption in single-organism RNA-seq that each cell harbors one copy of each gene, microbial communities can exhibit varying gene counts due to loss or duplication within strain-specific variations. This contributes to the occurrence of RNA zeroes, and an increase in gene copies correlates with heightened transcript abundance.

Consequently, community-level metatranscriptomics shows a strong correlation between DNA-level and RNA-level abundance of functions across samples. As a result, differences in metatranscriptomics abundance often stem from underlying gene copy number variations (metagenomic abundance), rather than actual differential expression or functional activity.

To counterbalance the pronounced influence of functional potential on community activity, a strategy is needed to gauge the relative expression of functions. This involves determining how much a function is over- or underexpressed in a metatranscriptome relative to the abundance of community genes encoding it. When genes can be attributed to specific species, a normalization approach involves adjusting transcript abundances within species. This method assumes uniform gene copy numbers within a species, akin to separate single-organism RNA-seq datasets. Alternatively, for communities with paired metatranscriptomics and metagenomics data, normalizing gene family RNA abundance by its DNA abundance can be applied. This approach is suitable for genes without known taxonomy and accounts for gene loss and duplication events (Y. Zhang et al., 2021).

6 Future directions

The science of environmental genomics continues to advance at a dizzying pace. Projects are producing ever-increasing volumes of data for a number of reasons: (1) as the scope of projects gets broader, more sites are sampled, more samples are collected per site, and multi-year data must be re-analyzed together as a whole to facilitate direct comparisons; (2) new DNA sequencing technologies means the depth of sequencing per sample has increased greatly, sometimes involving more than 100 million DNA sequence reads per sample in the case of metagenomics/metatranscriptomics data; and (3) reference databases are always getting bigger, increasing the time required to make taxonomic assignments. Analyzing these huge datasets within a reasonable timeframe will require advances both algorithmically and computationally. For example, NCBI Blast 2.15.0, released October 2023, included improvements that caused a significant speed-up in CEGA's workflow (Camacho & Madden, 2023)—a surprising achievement for a 33-year-old piece of software (Altschul et al., 1990)! On the computational side, CPUs continue to add more cores allowing computations to run in parallel. What would have been considered supercomputer cluster just a few decades ago is now available on a single chip: AMD's 7995WX CPU has 96 cores and is able to run 192 threads simultaneously (*Processor Specifications | AMD*, n.d.). Different computing architectures like graphical processing units (GPUs) and field

programmable gate arrays (FPGAs) can also allow certain mathematical operations to be performed much faster than normal CPUs, and can have applications within bioinformatics analyses (Nobile et al., 2017).

Turning raw DNA sequences into species identifications requires a robust reference database. Global scientific efforts are rapidly increasing the quantity and quality of reference DNA barcodes, but gaps still exist—particularly in areas of the world that lack the resources for advanced scientific studies, or areas of the world that are difficult or dangerous to study (e.g., the deepest depths of the ocean). Targeted efforts will likely be required to close these gaps.

The interpretation of environmental genomics data is still in its infancy. For the most part, analysts are just plugging species lists into existing ecological modeling frameworks, which does not take full advantage of the power that environmental genomics can bring to the table. Biological indices have historically built from small sets of species because it was difficult to measure more than that, but environmental genomics opens the possibility of measuring hundreds or even thousands of species easily and efficiently.

One significant drawback of eDNA data is that quantitative measurements are not yet robust, so the technique is limited to presence/absence measurements. However, this is an intense area of research and there have been some promising results suggesting that quantitative measurements may be possible in the near future.

While there are no formal standards yet established for the reporting of metabarcoding data, we suggest the following at minimum: sequence read length, QC criteria, read count per sample before and after quality filtering, software packages and parameter setting used in each step of the pipeline, OTU/ASV assignments, minimal read thresholds, and reference libraries used. These minimum reporting requirements will improve confidence in eDNA results and their interpretations and enhance the comparability between multiple studies and between eDNA practitioners. The wider integration of reporting standards will support eDNA data collation, data mining, and meta-analytical approaches for addressing larger-scale environmental questions.

Establishing standards for eDNA analysis and interpretation is key to facilitate its adoption by industry and regulatory agencies. Many of the debates about which analytical methods are “best” are largely academic since most of the popular choices are “good enough”—especially in light of the global biodiversity crisis and new regulations like the EU’s Corporate Sustainability Reporting Directive. But what is needed is for general agreement about which methods to use in which situations so that results are comparable across time and space. We hope this document serves as a step in that direction.

7 References

Abrego, N., Bässler, C., Christensen, M., & Heilmann-Clausen, J. (2022). Traits and phylogenies modulate the environmental responses of wood-inhabiting fungal communities across spatial scales. *Journal of Ecology*, *110*(4), 784–798. https://doi.org/10.1111/1365-2745.13839

Abrego, N., Crosier, B., Somervuo, P., Ivanova, N., Abrahamyan, A., Abdi, A., Hämäläinen, K., Junninen, K., Maunula, M., Purhonen, J., & Ovaskainen, O. (2020). Fungal communities decline with urbanization—More in air than in soil. *The ISME Journal*, *14*(11), 2806–2815. https://doi.org/10.1038/s41396-020-0732-1

Adamowicz, S. J. (2015). International Barcode of Life: Evolution of a global research community. *Genome*, *58*(5), 151–162. https://doi.org/10.1139/gen-2015-0094

Alberdi, A., Aizpurua, O., Bohmann, K., Gopalakrishnan, S., Lynggaard, C., Nielsen, M., & Gilbert, M. T. P. (2019). Promises and pitfalls of using high-throughput sequencing for diet analysis. *Molecular Ecology Resources*, *19*(2), 327–348. https://doi.org/10.1111/1755-0998.12960

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, *9*(1), 134–147. https://doi.org/10.1111/2041-210X.12849

Alberdi, A., Razgour, O., Aizpurua, O., Novella-Fernandez, R., Aihartza, J., Budinski, I., Garin, I., Ibáñez, C., Izagirre, E., Rebelo, H., Russo, D., Vlaschenko, A., Zhelyazkova, V., Zrnčić, V., & Gilbert, M. T. P. (2020). DNA metabarcoding and spatial modelling link diet diversification with distribution homogeneity in European bats. *Nature Communications*, *11*(1), 1154. https://doi.org/10.1038/s41467-020-14961-2

Alexander, J. B., Bunce, M., White, N., Wilkinson, S. P., Adam, A. A. S., Berry, T., Stat, M., Thomas, L., Newman, S. J., Dugal, L., & Richards, Z. T. (2020). Development of a multi-assay approach for monitoring coral diversity using eDNA metabarcoding. *Coral Reefs*, *39*(1), 159–171. https://doi.org/10.1007/s00338-019-01875-9

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), Article 11. https://doi.org/10.1038/nmeth.3103

Altschul, S. F. (2014). BLAST Algorithm. In *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470015902.a0005253.pub2

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Ames, S. K., Hysom, D. A., Gardner, S. N., Lloyd, G. S., Gokhale, M. B., & Allen, J. E. (2013). Scalable metagenomic taxonomy classification using a reference genome database. *Bioinformatics*, *29*(18), 2253–2260. https://doi.org/10.1093/bioinformatics/btt389

Andersen, M., Crist, T., Chase, J., Vellend, M., Inouye, B., Freestone, A., Sanders, N., Cornell, H., Comita, L., Davies, K., Harrison, S., Kraft, N., Stegen, J., & Swenson, N. (2011). Navigating the multiple meanings of β-diversity: A roadmap for the practicing ecologist. *Ecology Letters*, *14*, 19–28. https://doi.org/10.1111/j.1461-0248.2010.01552.x

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Andruszkiewicz, E. A., Starks, H. A., Chavez, F. P., Sassoubre, L. M., Block, B. A., & Boehm, A. B. (2017). Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. *PLOS ONE*, *12*(4), e0176343. https://doi.org/10.1371/journal.pone.0176343

Ansorge, R., Birolo, G., James, S. A., & Telatin, A. (2021). Dadaist2: A toolkit to automate and simplify statistical analysis and plotting of metabarcoding experiments. *International Journal of Molecular Sciences*, *22*(10), 5309.

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, *22*, 1–24.

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P., & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular Ecology Resources*, *17*(6), 1231–1242. https://doi.org/10.1111/1755-0998.12668

Austin, M. P. (2013). Inconsistencies between theory and methodology: A recurrent problem in ordination studies. *Journal of Vegetation Science*, *24*(2), 251–268. https://doi.org/10.1111/j.1654-1103.2012.01467.x

Aylagas, E., Borja, Á., Muxika, I., & Rodríguez-Ezpeleta, N. (2018). Adapting metabarcoding-based benthic biomonitoring into routine marine ecological status assessment networks. *Ecological Indicators*, *95*, 194–202. https://doi.org/10.1016/j.ecolind.2018.07.044

Aylagas, E., Borja, Á., & Rodríguez-Ezpeleta, N. (2014). Environmental Status Assessment Using DNA Metabarcoding: Towards a Genetics Based Marine Biotic Index (gAMBI). *PLoS ONE*, *9*(3), e90529. https://doi.org/10.1371/journal.pone.0090529

Aylagas, E., Borja, Á., Tangherlini, M., Dell'Anno, A., Corinaldesi, C., Michell, C. T., Irigoien, X., Danovaro, R., & Rodríguez-Ezpeleta, N. (2017). A bacterial community-based index to assess the ecological status of estuarine and coastal environments. *Marine Pollution Bulletin*, *114*(2), 679–688. https://doi.org/10.1016/j.marpolbul.2016.10.050

Banchi, E., Ametrano, C. G., Greco, S., Stanković, D., Muggia, L., & Pallavicini, A. (2020). PLANiTS: a curated sequence reference dataset for plant ITS DNA metabarcoding. *Database*, *2020*.

Banchi, E., Pallavicini, A., & Muggia, L. (2020). Relevance of plant and fungal DNA metabarcoding in aerobiology. *Aerobiologia*, *36*(1), 9–23. https://doi.org/10.1007/s10453-019-09574-2

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., & Stamatakis, A. (2019). EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Systematic Biology*, *68*(2), 365–369. https://doi.org/10.1093/sysbio/syy054

Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*(1), 1–17. https://doi.org/10.1007/s10592-015-0775-4

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., & Lodge, D. M. (2014a). Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology*, *48*(3), 1819–1827. https://doi.org/10.1021/es404734p

Barnes, M. A., Turner, C. R., Jerde, C. L., Renshaw, M. A., Chadderton, W. L., & Lodge, D. M. (2014b). Environmental Conditions Influence eDNA Persistence in Aquatic Systems. *Environmental Science & Technology*, *48*(3), 1819–1827. https://doi.org/10.1021/es404734p

Barroso-Bergadà, D., Pauvert, C., Vallance, J., Delière, L., Bohan, D. A., Buée, M., & Vacher, C. (2021). Microbial networks inferred from environmental DNA data for biomonitoring ecosystem change: Strengths and pitfalls. *Molecular Ecology Resources*, *21*(3), 762–780. https://doi.org/10.1111/1755-0998.13302

Barwell, L. J., Isaac, N. J. B., & Kunin, W. E. (2015). Measuring β-diversity with species abundance data. *Journal of Animal Ecology*, *84*(4), 1112–1122. https://doi.org/10.1111/1365-2656.12362

Bazinet, A. L., & Cummings, M. P. (2012). A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, *13*(1), 92. https://doi.org/10.1186/1471-2105-13-92

Beck, J., Holloway, J. D., & Schwanghart, W. (2013). Undersampling and the measurement of beta diversity. *Methods in Ecology and Evolution*, *4*(4), 370–382. https://doi.org/10.1111/2041-210x.12023

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., & Segata, N. (n.d.). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. *eLife*, *10*, e65088. https://doi.org/10.7554/eLife.65088

Bell, K. L., Burgess, K. S., Botsch, J. C., Dobbs, E. K., Read, T. D., & Brosi, B. J. (2019). Quantitative and qualitative assessment of pollen DNA metabarcoding using constructed species mixtures. *Molecular Ecology*, *28*(2), 431–455. https://doi.org/10.1111/mec.14840

Bennett, J. R., & Gilbert, B. (2016). Contrasting beta diversity among regions: How do classical and multivariate approaches compare?: Comparing regional differences in beta diversity. *Global Ecology and Biogeography*, *25*(3), 368–377. https://doi.org/10.1111/geb.12413

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Ostell, J., Pruitt, K. D., & Sayers, E. W. (2018). GenBank. *Nucleic Acids Research*, *46*(Database issue), D41.

Birk, S., Van Kouwen, L., & Willby, N. (2012). Harmonising the bioassessment of large rivers in the absence of near-natural reference conditions - a case study of the Danube River: Alternative benchmarking in large river bioassessment. *Freshwater Biology*, *57*(8), 1716–1732. https://doi.org/10.1111/j.1365-2427.2012.02831.x

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., Huttley, G. A., & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome*, *6*(1), 90. https://doi.org/10.1186/s40168-018-0470-z

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), Article 8. https://doi.org/10.1038/s41587-019-0209-9

Borja, A., Franco, J., & Pérez, V. (2000). A Marine Biotic Index to Establish the Ecological Quality of Soft-Bottom Benthos Within European Estuarine and Coastal Environments. *Marine Pollution Bulletin*, *40*(12), 1100–1114. https://doi.org/10.1016/S0025-326X(00)00061-8

Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods*, *6*(9), Article 9. https://doi.org/10.1038/nmeth.1358

Brantschen, J., Blackman, R. C., Walser, J.-C., & Altermatt, F. (2021). Environmental DNA gives comparable results to morphology-based indices of macroinvertebrates in a large-scale ecological assessment. *PLOS ONE*, *16*(9), e0257510. https://doi.org/10.1371/journal.pone.0257510

Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Elbrecht, V., Steinke, D., Ratnasingham, S., De Waard, J. R., Sones, J. E., Zakharov, E. V., & Hebert, P. D. N. (2019). Metabarcoding a diverse arthropod mock community. *Molecular Ecology Resources*, *19*(3), 711–727. https://doi.org/10.1111/1755-0998.13008

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), Article 4. https://doi.org/10.1038/s41592-021-01101-x

Buddle, C. M., Beguin, J., Bolduc, E., Mercado, A., Sackett, T. E., Selby, R. D., Varady-Szabo, H., & Zeran, R. M. (2005). The importance and use of taxon sampling curves for comparative biodiversity research with forest arthropod assemblages. *The Canadian Entomologist*, *137*(1), 120–127. https://doi.org/10.4039/n04-040

Bukin, Yu. S., Galachyants, Yu. P., Morozov, I. V., Bukin, S. V., Zakharenko, A. S., & Zemskaya, T. I. (2019). The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, *6*(1), 190007. https://doi.org/10.1038/sdata.2019.7

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*(1), 94. https://doi.org/10.1186/1471-2105-11-94

Burian, A., Mauvisseau, Q., Bulling, M., Domisch, S., Qian, S., & Sweet, M. (2021). Improving the reliability of eDNA data interpretation. *Molecular Ecology Resources*, *21*(5), 1422–1433. https://doi.org/10.1111/1755-0998.13367

Bush, A., Compson, Z., Rideout, N. K., Levenstein, B., Kattilakoski, M., Hajibabaei, M., Monk, W. A., Wright, M. T. G., & Baird, D. J. (2023). Replicate DNA metabarcoding can discriminate seasonal and spatial abundance shifts in river macroinvertebrate assemblages. *Molecular Ecology Resources*, *23*(6), 1275–1287. https://doi.org/10.1111/1755-0998.13794

Bush, A., Monk, W. A., Compson, Z. G., Peters, D. L., Porter, T. M., Shokralla, S., Wright, M. T. G., Hajibabaei, M., & Baird, D. J. (2020). DNA metabarcoding reveals metacommunity dynamics in a threatened boreal wetland wilderness. *Proceedings of the National Academy of Sciences*, *117*(15), 8539–8545. https://doi.org/10.1073/pnas.1918741117

Bushnell, B. (2014a). *BBDuk: Adapter/quality trimming and filtering*.

Bushnell, B. (2014b). *BBMap: A Fast, Accurate, Splice-Aware Aligner* (LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166

Bustin, S. A., Benes, V., Garson, J. A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M. W., Shipley, G. L., Vandesompele, J., & Wittwer, C. T. (2009). The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry*, *55*(4), 611–622. https://doi.org/10.1373/clinchem.2008.112797

Bylemans, J., Gleeson, D. M., Duncan, R. P., Hardy, C. M., & Furlan, E. M. (2019). A performance evaluation of targeted eDNA and eDNA metabarcoding analyses for freshwater fishes. *Environmental DNA*, *1*(4), 402–414. https://doi.org/10.1002/edn3.41

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Camacho, C., & Madden, T. (2023). BLAST+ Release Notes. In *BLAST® Help [Internet]*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK131777/

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., & Gordon, J. I. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336.

Capurso, G., Carroll, B., & Stewart, K. A. (2023). Transforming marine monitoring: Using eDNA metabarcoding to improve the monitoring of the Mediterranean Marine Protected Areas network. *Marine Policy*, *156*, 105807. https://doi.org/10.1016/j.marpol.2023.105807

Cardoso, P., Borges, P. A. V., & Veech, J. A. (2009). Testing the performance of beta diversity measures based on incidence data: The robustness to undersampling. *Diversity and Distributions*, *15*(6), 1081–1090. https://doi.org/10.1111/j.1472-4642.2009.00607.x

Carew, M. E., Pettigrove, V. J., Metzeling, L., & Hoffmann, A. A. (2013). Environmental monitoring using next generation sequencing: Rapid identification of macroinvertebrate bioindicator species. *Frontiers in Zoology*, *10*(1), 45. https://doi.org/10.1186/1742-9994-10-45

Carraro, L., Mächler, E., Wüthrich, R., & Altermatt, F. (2020). Environmental DNA allows upscaling spatial patterns of biodiversity in freshwater ecosystems. *Nature Communications*, *11*(1), 3585. https://doi.org/10.1038/s41467-020-17337-8

Chang, Q., Luan, Y., & Sun, F. (2011). Variance adjusted weighted UniFrac: A powerful beta diversity measure for comparing communities based on phylogeny. *BMC Bioinformatics*, *12*(1), 118. https://doi.org/10.1186/1471-2105-12-118

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T.-J. (2004). A new statistical approach for assessing similarity of species composition with incidence and abundance data: A new statistical approach for assessing similarity. *Ecology Letters*, *8*(2), 148–159. https://doi.org/10.1111/j.1461-0248.2004.00707.x

Chao, A., & Chiu, C. (2016). Bridging the variance and diversity decomposition approaches to beta diversity via similarity and differentiation measures. *Methods in Ecology and Evolution*, *7*(8), 919–928. https://doi.org/10.1111/2041-210X.12551

Chao, A., & Jost, L. (2012). Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology*, *93*(12), 2533–2547. https://doi.org/10.1890/11-1952.1

Chao, A., Thorn, S., Chiu, C., Moyes, F., Hu, K., Chazdon, R. L., Wu, J., Magnago, L. F. S., Dornelas, M., Zelený, D., Colwell, R. K., & Magurran, A. E. (2023). Rarefaction and extrapolation with beta diversity under a framework of Hill numbers: The iNEXT . BETA3D standardization. *Ecological Monographs*, e1588. https://doi.org/10.1002/ecm.1588

Chase, J. M., McGill, B. J., Thompson, P. L., Antão, L. H., Bates, A. E., Blowes, S. A., Dornelas, M., Gonzalez, A., Magurran, A. E., Supp, S. R., Winter, M., Bjorkman, A. D., Bruelheide, H., Byrnes, J. E. K., Cabral, J. S., Elahi, R., Gomez, C., Guzman, H. M., Isbell, F., … O'Connor, M. (2019). Species richness change across spatial scales. *Oikos*, *128*(8), 1079–1091. https://doi.org/10.1111/oik.05968

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: A toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, *36*(6), 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: An accelerated MegaBLAST search tool. *Nucleic Acids Research*, *43*(16), 7762–7768. https://doi.org/10.1093/nar/gkv784

Chiarello, M., McCauley, M., Villéger, S., & Jackson, C. R. (2022). Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. *PLoS One*, *17*(2), e0264443.

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., & Eichler, E. E. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, *10*(6), 563–569.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., & Schatz, M. C. (2016). Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing. *Nature Methods*, *13*(12), 1050–1054. https://doi.org/10.1038/nmeth.4035

Chiu, C.-H., & Chao, A. (2016). Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, *4*, e1634. https://doi.org/10.7717/peerj.1634

Clare, E. L., Fazekas, A. J., Ivanova, N. V., Floyd, R. M., Hebert, P. D. N., Adams, A. M., Nagel, J., Girton, R., Newmaster, S. G., & Fenton, M. B. (2019). Approaches to integrating genetic data into ecological networks. *Molecular Ecology*, *28*(2), 503–519. https://doi.org/10.1111/mec.14941

Cobo-Díaz, J. F., Baroncelli, R., Le Floch, G., & Picot, A. (2019). Combined Metabarcoding and Co-occurrence Network Analysis to Profile the Bacterial, Fungal and Fusarium Communities and Their Interactions in Maize Stalks. *Frontiers in Microbiology*, *10*, 261. https://doi.org/10.3389/fmicb.2019.00261

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., & Tiedje, J. M. (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, *37*(Database), D141–D145. https://doi.org/10.1093/nar/gkn879

Collins, R. A., Trauzzi, G., Maltby, K. M., Gibson, T. I., Ratcliffe, F. C., Hallam, J., Rainbird, S., Maclaine, J., Henderson, P. A., & Sims, D. W. (2021). Meta-Fish-Lib: A generalised, dynamic DNA reference library pipeline for metabarcoding of fishes. *Journal of Fish Biology*, *99*(4), 1446–1454.

Collins, R. A., Wangensteen, O. S., O'Gorman, E. J., Mariani, S., Sims, D. W., & Genner, M. J. (2018). Persistence of environmental DNA in marine systems. *Communications Biology*, *1*(185), 1–11. https://doi.org/10.1038/s42003-018-0192-6

Colwell, R., & Coddington, J. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *345*(1311), 101–118. https://doi.org/10.1098/rstb.1994.0091

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., & Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, *5*(1), 3–21. https://doi.org/10.1093/jpe/rtr044

Compson, Z. G., McClenaghan, B., Singer, G. A. C., Fahner, N. A., & Hajibabaei, M. (2020). Metabarcoding From Microbes to Mammals: Comprehensive Bioassessment on a Global Scale. *Frontiers in Ecology and Evolution*, *8*, 581835. https://doi.org/10.3389/fevo.2020.581835

Compson, Z. G., Monk, W. A., Curry, C. J., Gravel, D., Bush, A., Baker, C. J. O., Al Manir, M. S., Riazanov, A., Hajibabaei, M., Shokralla, S., Gibson, J. F., Stefani, S., Wright, M. T. G., & Baird, D. J. (2018). Linking DNA Metabarcoding and Text Mining to Create Network-Based Biomonitoring Tools: A Case Study on Boreal Wetland Macroinvertebrate Communities. In *Advances in Ecological Research* (Vol. 59, pp. 33–74). Elsevier. https://doi.org/10.1016/bs.aecr.2018.09.001

Cordier, T. (2020). Bacterial communities' taxonomic and functional turnovers both accurately predict marine benthic ecological quality status. *Environmental DNA*, *2*(2), 175–183. https://doi.org/10.1002/edn3.55

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2021). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*(13), 2937–2958. https://doi.org/10.1111/mec.15472

Cordier, T., Esling, P., Lejzerowicz, F., Visco, J., Ouadahi, A., Martins, C., Cedhagen, T., & Pawlowski, J. (2017). Predicting the Ecological Quality Status of Marine Environments from eDNA Metabarcoding Data Using Supervised Machine Learning. *Environmental Science & Technology*, *51*(16), 9118–9126. https://doi.org/10.1021/acs.est.7b01518

Cordier, T., Frontalini, F., Cermakova, K., Apothéloz-Perret-Gentil, L., Treglia, M., Scantamburlo, E., Bonamin, V., & Pawlowski, J. (2019). Multi-marker eDNA metabarcoding survey to assess the environmental impact of three offshore gas platforms in the North Adriatic Sea (Italy). *Marine Environmental Research*, *146*, 24–34. https://doi.org/10.1016/j.marenvres.2018.12.009

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., & Arnaud-Haond, S. (2015). Metabarcoding Is Powerful yet Still Blind: A Comparative Analysis of Morphological and Molecular Surveys of Seagrass Communities. *PLOS ONE*, *10*(2), e0117562. https://doi.org/10.1371/journal.pone.0117562

Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution*, *29*(10), 566–571. https://doi.org/10.1016/j.tree.2014.08.001

D'Alessandro, S., & Mariani, S. (2021). Sifting environmental DNA metabarcoding data sets for rapid reconstruction of marine food webs. *Fish and Fisheries*, *22*(4), 822–833. https://doi.org/10.1111/faf.12553

Daly, A., Baetens, J., & De Baets, B. (2018). Ecological Diversity: Measuring the Unmeasurable. *Mathematics*, *6*(7), 119. https://doi.org/10.3390/math6070119

Darling, J. A., Jerde, C. L., & Sepulveda, A. J. (2021). What do you mean by false positive? *Environmental DNA*, *3*(5), 879–883. https://doi.org/10.1002/edn3.194

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 226. https://doi.org/10.1186/s40168-018-0605-2

De Coster, W., & Rademakers, R. (2023). NanoPack2: Population-scale evaluation of long-read sequencing data. *Bioinformatics*, *39*(5), btad311.

Deagle, B. E., Thomas, A. C., Shaffer, A. K., Trites, A. W., & Jarman, S. N. (2013). Quantifying sequence proportions in a DNA -based diet study using Ion Torrent amplicon sequencing: Which counts count? *Molecular Ecology Resources*, *13*(4), 620–633. https://doi.org/10.1111/1755-0998.12103

Dejean, T., Valentini, A., Duparc, A., Pellier-Cuit, S., Pompanon, F., Taberlet, P., & Miaud, C. (2011). Persistence of environmental DNA in freshwater ecosystems. *PLoS ONE*, *6*(8), e23398. https://doi.org/10.1371/journal.pone.0023398

Di Muri, C., Lawson Handley, L., Bean, C. W., Li, J., Peirson, G., Sellers, G. S., Walsh, K., Watson, H. V., Winfield, I. J., & Hänfling, B. (2020). Read counts from environmental DNA (eDNA) metabarcoding reflect fish abundance and biomass in drained ponds. *Metabarcoding and Metagenomics*, *4*, e56959. https://doi.org/10.3897/mbmg.4.56959

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319.

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., & Banfield, J. F. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biology*, *10*(8), R85. https://doi.org/10.1186/gb-2009-10-8-r85

Doi, H., Fukaya, K., Oka, S., Sato, K., Kondoh, M., & Miya, M. (2019). Evaluation of detection probabilities at the water-filtering and initial PCR steps in environmental DNA metabarcoding using a multispecies site occupancy model. *Scientific Reports*, *9*(3581), 1–8. https://doi.org/10.1038/s41598-019-40233-1

Doi, H., Inui, R., Matsuoka, S., Akamatsu, Y., Goto, M., & Kono, T. (2021). Estimation of biodiversity metrics by environmental DNA metabarcoding compared with visual and capture surveys of river fish communities. *Freshwater Biology*, *66*(7), 1257–1266. https://doi.org/10.1111/fwb.13714

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., & Langille, M. G. I. (2020). PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, *38*(6), Article 6. https://doi.org/10.1038/s41587-020-0548-6

Drake, L. E., Cuff, J. P., Young, R. E., Marchbank, A., Chadwick, E. A., & Symondson, W. O. C. (2022). An assessment of minimum sequence copy thresholds for identifying and reducing the prevalence of artefacts in dietary metabarcoding data. *Methods in Ecology and Evolution*, *13*(3), 694–710. https://doi.org/10.1111/2041-210X.13780

Dubois, B., Debode, F., Hautier, L., Hulin, J., Martin, G. S., Delvaux, A., Janssen, E., & Mingeot, D. (2022). A detailed workflow to develop QIIME2-formatted reference databases for taxonomic analysis of DNA metabarcoding data. *BMC Genomic Data*, *23*(1), 53.

Dutilh, B. E., Huynen, M. A., & Strous, M. (2009). Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics (Oxford, England)*, *25*(21), 2878–2881. https://doi.org/10.1093/bioinformatics/btp377

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461.

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed-and open-reference OTUs. *PeerJ*, *5*, e3889.

Elbrecht, V., & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol. *PLOS ONE*, *10*(7), e0130324. https://doi.org/10.1371/journal.pone.0130324

Ershova, E. A., Wangensteen, O. S., Descoteaux, R., Barth-Jensen, C., & Præbel, K. (2021). Metabarcoding as a quantitative tool for estimating biodiversity and relative biomass of marine zooplankton. *ICES Journal of Marine Science*, *78*(9), 3342–3355. https://doi.org/10.1093/icesjms/fsab171

Evans, D. M., Kitson, J. J. N., Lunt, D. H., Straw, N. A., & Pocock, M. J. O. (2016). Merging DNA metabarcoding and ecological network analysis to understand and build resilient terrestrial ecosystems. *Functional Ecology*, *30*(12), 1904–1916. https://doi.org/10.1111/1365-2435.12659

Evans, N. T., Li, Y., Renshaw, M. A., Olds, B. P., Deiner, K., Turner, C. R., Jerde, C. L., Lodge, D. M., Lamberti, G. A., & Pfrender, M. E. (2017). Fish community assessment with eDNA metabarcoding: Effects of sampling design and bioinformatic filtering. *Canadian Journal of Fisheries and Aquatic Sciences*, *74*(9), 1362–1374. https://doi.org/10.1139/cjfas-2016-0306

Evans, N. T., Olds, B. P., Renshaw, M. A., Turner, C. R., Li, Y., Jerde, C. L., Mahon, A. R., Pfrender, M. E., Lamberti, G. A., & Lodge, D. M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 29–41. https://doi.org/10.1111/1755-0998.12433

Fais, M., Bellisario, B., Duarte, S., Vieira, P. E., Sousa, R., Canchaya, C., & Costa, F. O. (2020). Meiofauna metabarcoding in Lima estuary (Portugal) suggests high taxon replacement within a background of network stability. *Regional Studies in Marine Science*, *38*, 101341. https://doi.org/10.1016/j.rsma.2020.101341

Fernández, S., Rodríguez-Martínez, S., Martínez, J. L., Garcia-Vazquez, E., & Ardura, A. (2019). How can eDNA contribute in riverine macroinvertebrate assessment? A metabarcoding approach in the Nalón River (Asturias, Northern Spain). *Environmental DNA*, *1*(4), 385–401. https://doi.org/10.1002/edn3.40

Ficetola, G. F., Pansu, J., Bonin, A., Coissac, E., Giguet-Covex, C., De Barba, M., Gielly, L., Lopes, C. M., Boyer, F., Pompanon, F., Rayé, G., & Taberlet, P. (2015). Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. *Molecular Ecology Resources*, *15*(3), 543–556. https://doi.org/10.1111/1755-0998.12338

Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding? *Molecular Ecology Resources*, *16*(3), 604–607. https://doi.org/10.1111/1755-0998.12508

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., Ashburner, M., Axelrod, N., Baldauf, S., Ballard, S., Boore, J., Cochrane, G., Cole, J., Dawyndt, P., De Vos, P., … Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, *26*(5), 541–547. https://doi.org/10.1038/nbt1360

Foucher, A., Evrard, O., Ficetola, G. F., Gielly, L., Poulain, J., Giguet-Covex, C., Laceby, J. P., Salvador-Blanes, S., Cerdan, O., & Poulenard, J. (2020). Persistence of environmental DNA in cultivated soils: Implication of this memory effect for reconstructing the dynamics of land use and cover changes. *Scientific Reports*, *10*(1), 10502. https://doi.org/10.1038/s41598-020-67452-1

Foulon, J., Zappelini, C., Durand, A., Valot, B., Girardclos, O., Blaudez, D., & Chalot, M. (2016). Environmental metabarcoding reveals contrasting microbial communities at two poplar phytomanagement sites. *Science of The Total Environment*, *571*, 1230–1240. https://doi.org/10.1016/j.scitotenv.2016.07.151

Frankenfeld, C. L., Hullar, M. A. J., Maskarinec, G., Monroe, K. R., Shepherd, J. A., Franke, A. A., Randolph, T. W., Wilkens, L. R., Boushey, C. J., Le Marchand, L., Lim, U., & Lampe, J. W. (2022). The Gut Microbiome Is Associated with Circulating Dietary Biomarkers of Fruit and Vegetable Intake in a Multiethnic Cohort. *Journal of the Academy of Nutrition and Dietetics*, *122*(1), 78–98. https://doi.org/10.1016/j.jand.2021.05.023

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., & Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, *15*(11), Article 11. https://doi.org/10.1038/s41592-018-0176-y

Fremier, A. K., Strickler, K. M., Parzych, J., Powers, S., & Goldberg, C. S. (2019). Stream Transport and Retention of Environmental DNA Pulse Releases in Relation to Hydrogeomorphic Scaling Factors. *Environmental Science & Technology*, *53*(12), 6640–6649. https://doi.org/10.1021/acs.est.8b06829

Fujii, K., Doi, H., Matsuoka, S., Nagano, M., Sato, H., & Yamanaka, H. (2019). Environmental DNA metabarcoding for fish community analysis in backwater lakes: A comparison of capture methods. *PLOS ONE*, *14*(1), e0210357. https://doi.org/10.1371/journal.pone.0210357

Fukasawa, Y., Matsukura, K., Stephan, J. G., Makoto, K., Suzuki, S. N., Kominami, Y., Takagi, M., Tanaka, N., Takemoto, S., Kinuura, H., Okano, K., Song, Z., Jomura, M., Kadowaki, K., Yamashita, S., & Ushio, M. (2022). Patterns of community composition and diversity in latent fungi of living Quercus serrata trunks across a range of oak wilt prevalence and climate variables in Japan. *Fungal Ecology*, *59*, 101095. https://doi.org/10.1016/j.funeco.2021.101095

Galloway-Peña, J., & Hanson, B. (2020). Tools for Analysis of the Microbiome. *Digestive Diseases and Sciences*, *65*(3), 674–685. https://doi.org/10.1007/s10620-020-06091-y

Garcia-Vazquez, E., Georges, O., Fernandez, S., & Ardura, A. (2021). eDNA metabarcoding of small plankton samples to detect fish larvae and their preys from Atlantic and Pacific waters. *Scientific Reports*, *11*(1), 7224. https://doi.org/10.1038/s41598-021-86731-z

Gaston, K. J., Blackburn, T. M., Greenwood, J. J. D., Gregory, R. D., Quinn, R. M., & Lawton, J. H. (2000). Abundance–occupancy relationships. *Journal of Applied Ecology*, *37*(s1), 39–59. https://doi.org/10.1046/j.1365-2664.2000.00485.x

Glassman, S. I., & Martiny, J. B. H. (2018). Broadscale Ecological Patterns Are Robust to Use of Exact Sequence Variants versus Operational Taxonomic Units. *mSphere*, *3*(4), 10.1128/msphere.00148-18. https://doi.org/10.1128/msphere.00148-18

Gold, Z., Choi, E., Kacev, D., Frable, B., Burton, R., Goodwin, K., Thompson, A., & Barber, P. (2020). FishCARD: Fish 12S California current specific reference database for enhanced metabarcoding efforts. *Authorea Preprints*.

Gold, Z., Kelly, R. P., Shelton, A. O., Thompson, A. R., Goodwin, K. D., Gallego, R., Parsons, K. M., Thompson, L. R., Kacev, D., & Barber, P. H. (2023). Archived DNA reveals marine HEATWAVE-ASSOCIATED shifts in fish assemblages. *Environmental DNA*, edn3.400. https://doi.org/10.1002/edn3.400

Gold, Z., Shelton, A. O., Casendino, H. R., Duprey, J., Gallego, R., Van Cise, A., Fisher, M., Jensen, A. J., D'Agnese, E., Andruszkiewicz Allan, E., Ramón-Laca, A., Garber-Yonts, M., Labare, M., Parsons, K. M., & Kelly, R. P. (2023). Signal and noise in metabarcoding data. *PLOS ONE*, *18*(5), e0285674. https://doi.org/10.1371/journal.pone.0285674

Goldberg, C. S., Turner, C. R., Deiner, K., Klymus, K. E., Thomsen, P. F., Murphy, M. A., Spear, S. F., McKee, A., Oyler-McCance, S. J., Cornman, R. S., Laramie, M. B., Mahon, A. R., Lance, R. F., Pilliod, D. S., Strickler, K. M., Waits, L. P., Fremier, A. K., Takahara, T., Herder, J. E., & Taberlet, P. (2016). Critical considerations for

the application of environmental DNA methods to detect aquatic species. *Methods in Ecology and Evolution*, *7*(11), 1299–1307. https://doi.org/10.1111/2041-210X.12595

González, A., Dubut, V., Corse, E., Mekdad, R., Dechatre, T., Castet, U., Hebert, R., & Meglécz, E. (2023). VTAM: A robust pipeline for validating metabarcoding data using controls. *Computational and Structural Biotechnology Journal*, *21*, 1151–1156. https://doi.org/10.1016/j.csbj.2023.01.034

Gotelli, N. J., & Colwell, R. K. (n.d.). *Estimating species richness*.

Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, *4*(4), 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x

Graham, N. R., Gillespie, R. G., & Krehenwinkel, H. (2021). Towards eradicating the nuisance of numts and noise in molecular biodiversity assessment. *Molecular Ecology Resources*, *21*(6), 1755–1758. https://doi.org/10.1111/1755-0998.13414

Guardiola, M., Wangensteen, O. S., Taberlet, P., Coissac, E., Uriz, M. J., & Turon, X. (2016). Spatio-temporal monitoring of deep-sea communities using metabarcoding of sediment DNA and RNA. *PeerJ*, *4*, e2807. https://doi.org/10.7717/peerj.2807

Guillera-Arroita, G., Lahoz-Monfort, J. J., van Rooyen, A. R., Weeks, A. R., & Tingley, R. (2017). Dealing with false-positive and false-negative errors about species occurrence at multiple levels. *Methods in Ecology and Evolution*, *8*(9), 1081–1091. https://doi.org/10.1111/2041-210X.12743

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hajibabaei, M., Porter, T. M., Wright, M., & Rudar, J. (2019). COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PloS One*, *14*(9), e0220953. https://doi.org/10.1371/journal.pone.0220953

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., & Baird, D. J. (2011). Environmental barcoding: A next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS ONE*, *6*(4), e17497. https://doi.org/10.1371/journal.pone.0017497

Hajibabaei, M., Singer, G. A. C., Hebert, P. D. N., & Hickey, D. A. (2007). DNA barcoding: How it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics : TIG*, *23*(4), 167–172. https://doi.org/10.1016/j.tig.2007.02.001

Hakimzadeh, A., Abdala Asbun, A., Albanese, D., Bernard, M., Buchner, D., Callahan, B., Caporaso, J. G., Curd, E., Djemiel, C., & Brandström Durling, M. (2023). A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. *Molecular Ecology Resources*.

Hamer, A. J., Schmera, D., & Mahony, M. J. (2021). Multi-species occupancy modeling provides novel insights into amphibian metacommunity structure and wetland restoration. *Ecological Applications*, *31*(4). https://doi.org/10.1002/eap.2293

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249. https://doi.org/10.1016/S1074-5521(98)90108-9

He, X., Gilmore, S., Sutherland, T., Hajibabaei, M., Miller, K., Westfall, K., Pawlowski, J., & Abbott, C. (2020). *Biotic signals associated with benthic impacts of salmon farms from eDNA metabarcoding of sediments* [Preprint]. Preprints. https://doi.org/10.22541/au.159986543.39478542

He, X., Jeffery, N. W., Stanley, R. R. E., Hamilton, L. C., Rubidge, E. M., & Abbott, C. L. (2023). eDNA metabarcoding enriches traditional trawl survey data for monitoring biodiversity in the marine environment. *ICES Journal of Marine Science*, *80*(5), 1529–1538. https://doi.org/10.1093/icesjms/fsad083

He, X., Stanley, R. R. E., Rubidge, E. M., Jeffery, N. W., Hamilton, L. C., Westfall, K. M., Gilmore, S. R., Roux, L.-M. D., Gale, K. S. P., Heaslip, S. G., Steeves, R., & Abbott, C. L. (2022). Fish community surveys in eelgrass

beds using both eDNA metabarcoding and seining: Implications for biodiversity monitoring in the coastal zone. *Canadian Journal of Fisheries and Aquatic Sciences*, 1–12. https://doi.org/10.1139/cjfas-2021-0215

Hestetun, J. T., Bye-Ingebrigtsen, E., Nilsson, R. H., Glover, A. G., Johansen, P.-O., & Dahlgren, T. G. (2020). Significant taxon sampling gaps in DNA databases limit the operational use of marine macrofauna metabarcoding. *Marine Biodiversity*, *50*(5), 70. https://doi.org/10.1007/s12526-020-01093-5

Hestetun, J. T., Lanzén, A., & Dahlgren, T. G. (2021). Grab what you can—An evaluation of spatial replication to decrease heterogeneity in sediment eDNA metabarcoding. *PeerJ*, *9*, e11619. https://doi.org/10.7717/peerj.11619

Hleap, J. S., Littlefair, J. E., Steinke, D., Hebert, P. D., & Cristescu, M. E. (2021). Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. *Molecular Ecology Resources*, *21*(7), 2190–2203.

Holt, E., & Miller, S. (2010). Bioindicators: Using organisms to measure environmental impacts. *Nature Education Knowledge*, *3*(10).

Horn, S., De La Vega, C., Asmus, R., Schwemmer, P., Enners, L., Garthe, S., Haslob, H., Binder, K., & Asmus, H. (2019). Impact of birds on intertidal food webs assessed with ecological network analysis. *Estuarine, Coastal and Shelf Science*, *219*, 107–119. https://doi.org/10.1016/j.ecss.2019.01.023

Huang, X., Wang, J., Aluru, S., Yang, S.-P., & Hillier, L. (2003). PCAP: A Whole-Genome Assembly Program. *Genome Research*, *13*(9), 2164–2170. https://doi.org/10.1101/gr.1390403

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Huo, S., Li, X., Xi, B., Zhang, H., Ma, C., & He, Z. (2020). Combining morphological and metabarcoding approaches reveals the freshwater eukaryotic phytoplankton community. *Environmental Sciences Europe*, *32*(1), 37. https://doi.org/10.1186/s12302-020-00321-w

Huson, D. H., Mitra, S., Ruscheweyh, H.-J., Weber, N., & Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, *21*(9), 1552–1560. https://doi.org/10.1101/gr.120618.111

Iwai, S., Weinmaier, T., Schmidt, B. L., Albertson, D. G., Poloso, N. J., Dabbagh, K., & DeSantis, T. Z. (2016). Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes. *PloS One*, *11*(11), e0166104. https://doi.org/10.1371/journal.pone.0166104

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*, 1–11.

Jerde, C. L. (2021). Can we manage fisheries with the inherent uncertainty from eDNA? *Journal of Fish Biology*, *98*(2), 341–353. https://doi.org/10.1111/jfb.14218

Jeunen, G., Knapp, M., Spencer, H. G., Lamare, M. D., Taylor, H. R., Stat, M., Bunce, M., & Gemmell, N. J. (2019). Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement. *Molecular Ecology Resources*, *19*(2), 426–438. https://doi.org/10.1111/1755-0998.12982

Jeunen, G.-J., Dowle, E., Edgecombe, J., von Ammon, U., Gemmell, N. J., & Cross, H. (2023). crabs—A software program to generate curated reference databases for metabarcoding sequencing data. *Molecular Ecology Resources*, *23*(3), 725–738. https://doi.org/10.1111/1755-0998.13741

Johansen, J., Plichta, D. R., Nissen, J. N., Jespersen, M. L., Shah, S. A., Deng, L., Stokholm, J., Bisgaard, H., Nielsen, D. S., Sørensen, S. J., & Rasmussen, S. (2022). Genome binning of viral entities from bulk metagenomics data. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-28581-5

Joseph, C., Faiq, M. E., Li, Z., & Chen, G. (2022). Persistence and degradation dynamics of eDNA affected by environmental factors in aquatic ecosystems. *Hydrobiologia*, *849*(19), 4119–4133. https://doi.org/10.1007/s10750-022-04959-w

Jost, L. (2006). *Entropy and diversity*. *Oikos*, *113*(2), 363–375. https://doi.org/10.1111/j.2006.0030-1299.14714.x

Jupke, J. F., & Schäfer, R. B. (2020). Should ecologists prefer model- over distance-based multivariate methods? *Ecology and Evolution*, *10*(5), 2417–2435. https://doi.org/10.1002/ece3.6059

Kačergytė, I., Knape, J., Żmihorski, M., Arlt, D., & Pärt, T. (2023). Community associations of birds with amphibians and fish in wetlands created for biodiversity. *Biological Conservation*, *282*, 110031. https://doi.org/10.1016/j.biocon.2023.110031

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, e7359. https://doi.org/10.7717/peerj.7359

Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W. T., Wolfe, A. J., Braun, J., & McWeeney, S. (2019). Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems*, *4*(4), e00290-19. https://doi.org/10.1128/mSystems.00290-19

Keck, F., & Altermatt, F. (2023). Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. *Molecular Ecology Resources*, *23*(2), 511–518.

Keck, F., Blackman, R. C., Bossart, R., Brantschen, J., Couton, M., Hürlemann, S., Kirschner, D., Locher, N., Zhang, H., & Altermatt, F. (2022). Meta-analysis shows both congruence and complementarity of DNA and eDNA metabarcoding to traditional methods for biological community assessment. *Molecular Ecology*, *31*(6), 1820–1835. https://doi.org/10.1111/mec.16364

Keeley, N., Wood, S. A., & Pochon, X. (2018). Development and preliminary validation of a multi-trophic metabarcoding biotic index for monitoring benthic organic enrichment. *Ecological Indicators*, *85*, 1044–1057. https://doi.org/10.1016/j.ecolind.2017.11.014

Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., & Ankenbrand, M. J. (2020). BCdatabaser: On-the-fly reference database creation for (meta-) barcoding. *Bioinformatics*, *36*(8), 2630–2631.

Kelly, M. G., Juggins, S., Mann, D. G., Sato, S., Glover, R., Boonham, N., Sapp, M., Lewis, E., Hany, U., Kille, P., Jones, T., & Walsh, K. (2020). Development of a novel metric for evaluating diatom assemblages in rivers using DNA metabarcoding. *Ecological Indicators*, *118*, 106725. https://doi.org/10.1016/j.ecolind.2020.106725

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: Rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729. https://doi.org/10.1101/gr.210641.116

Kimble, M., Allers, S., Campbell, K., Chen, C., Jackson, L. M., King, B. L., Silverbrand, S., York, G., & Beard, K. (2022). medna-metadata: An open-source data management system for tracking environmental DNA samples and metadata. *Bioinformatics*, *38*(19), 4589–4597. https://doi.org/10.1093/bioinformatics/btac556

Klymus, K. E., Marshall, N. T., & Stepien, C. A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLOS ONE*, *12*(5), e0177643. https://doi.org/10.1371/journal.pone.0177643

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., & Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, *8*(9), 761–763. https://doi.org/10.1038/nmeth.1650

Koleff, P., Gaston, K., & Lennon, J. (2003). Measuring beta diversity for presence –absence data. *Journal of Animal Ecology*, *72*, 367–382.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Köster, J., & Rahmann, S. (2012). Snakemake—A scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522.

Krah, F., & March-Salas, M. (2022). eDNA metabarcoding reveals high soil fungal diversity and variation in community composition among Spanish cliffs. *Ecology and Evolution*, *12*(12), e9594. https://doi.org/10.1002/ece3.9594

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, *7*(1), 17668. https://doi.org/10.1038/s41598-017-17333-x

Kuntke, F., De Jonge, N., Hesselsøe, M., & Lund Nielsen, J. (2020). Stream water quality assessment by metabarcoding of invertebrates. *Ecological Indicators*, *111*, 105982. https://doi.org/10.1016/j.ecolind.2019.105982

Lacoursière-Roussel, A., Howland, K., Normandeau, E., Grey, E. K., Archambault, P., Deiner, K., Lodge, D. M., Hernandez, C., Leduc, N., & Bernatchez, L. (2018). eDNA metabarcoding as a new surveillance approach for coastal Arctic biodiversity. *Ecology and Evolution*, *8*(16), 7763–7777. https://doi.org/10.1002/ece3.4213

Lacoursière-Roussel, A., Rosabal, M., & Bernatchez, L. (2016). Estimating fish abundance and biomass from eDNA concentrations: Variability among capture methods and environmental conditions. *Molecular Ecology Resources*, *16*(6), 1401–1414. https://doi.org/10.1111/1755-0998.12522

Lahoz-Monfort, J. J., Guillera-Arroita, G., & Tingley, R. (2016). Statistical approaches to account for false-positive errors in environmental DNA samples. *Molecular Ecology Resources*, *16*(3), 673–685. https://doi.org/10.1111/1755-0998.12486

Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, *28*(2), 420–430. https://doi.org/10.1111/mec.14920

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C., Burkepile, D. E., Vega Thurber, R. L., Knight, R., Beiko, R. G., & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, *31*(9), Article 9. https://doi.org/10.1038/nbt.2676

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lanzén, A., Dahlgren, T. G., Bagi, A., & Hestetun, J. T. (2021). Benthic eDNA metabarcoding provides accurate assessments of impact from oil extraction, and ecological insights. *Ecological Indicators*, *130*, 108064. https://doi.org/10.1016/j.ecolind.2021.108064

Lanzén, A., Mendibil, I., Borja, Á., & Alonso-Sáez, L. (2021). A microbial *mandala* for environmental monitoring: Predicting multiple impacts on estuarine prokaryote communities of the Bay of Biscay. *Molecular Ecology*, *30*(13), 2969–2987. https://doi.org/10.1111/mec.15489

Laroche, O., Wood, S. A., Tremblay, L. A., Ellis, J. I., Lejzerowicz, F., Pawlowski, J., Lear, G., Atalah, J., & Pochon, X. (2016). First evaluation of foraminiferal metabarcoding for monitoring environmental impact from an offshore oil drilling site. *Marine Environmental Research*, *120*, 225–235. https://doi.org/10.1016/j.marenvres.2016.08.009

Laroche, O., Wood, S. A., Tremblay, L. A., Lear, G., Ellis, J. I., & Pochon, X. (2017). Metabarcoding monitoring analysis: The pros and cons of using co-extracted environmental DNA and RNA data to assess offshore oil production impacts on benthic communities. *PeerJ*, *5*, e3347. https://doi.org/10.7717/peerj.3347

Larsen, B. B., Miller, E. C., Rhodes, M. K., & Wiens, J. J. (2017). Inordinate fondness multiplied and redistributed: The number of species on earth and the new pie of life. *The Quarterly Review of Biology*, *92*(3), 229–265.

Leduc, N., Lacoursière-Roussel, A., Howland, K. L., Archambault, P., Sevellec, M., Normandeau, E., Dispas, A., Winkler, G., McKindsey, C. W., Simard, N., & Bernatchez, L. (2019). Comparing eDNA metabarcoding and species collection for documenting Arctic metazoan biodiversity. *Environmental DNA*, *1*(4), 342–358. https://doi.org/10.1002/edn3.35

Leite, M. F. A., & Kuramae, E. E. (2020). You must choose, but choose wisely: Model-based approaches for microbial community analysis. *Soil Biology and Biochemistry*, *151*, 108042. https://doi.org/10.1016/j.soilbio.2020.108042

Lejzerowicz, F., Gooday, A. J., Barrenechea Angeles, I., Cordier, T., Morard, R., Apothéloz-Perret-Gentil, L., Lins, L., Menot, L., Brandt, A., Levin, L. A., Martinez Arbizu, P., Smith, C. R., & Pawlowski, J. (2021). Eukaryotic Biodiversity and Spatial Patterns in the Clarion-Clipperton Zone and Other Abyssal Regions: Insights From Sediment DNA and RNA Metabarcoding. *Frontiers in Marine Science*, *8*, 671033. https://doi.org/10.3389/fmars.2021.671033

Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656.

Leray, M., Knowlton, N., & Machida, R. J. (2022). MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. *Environmental DNA*, *4*(4), 894–907. https://doi.org/10.1002/edn3.303

Li, H. (2015). BFC: Correcting Illumina sequencing errors. *Bioinformatics*, *31*(17), 2885–2887. https://doi.org/10.1093/bioinformatics/btv290

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, W., Hou, X., Xu, C., Qin, M., Wang, S., Wei, L., Wang, Y., Liu, X., & Li, Y. (2021). Validating eDNA measurements of the richness and abundance of anurans at a large scale. *Journal of Animal Ecology*, *90*(6), 1466–1479. https://doi.org/10.1111/1365-2656.13468

Li, Y., Evans, N. T., Renshaw, M. A., Jerde, C. L., Olds, B. P., Shogren, A. J., Deiner, K., Lodge, D. M., Lamberti, G. A., & Pfrender, M. E. (2018). Estimating fish alpha- and beta-diversity along a small stream with environmental DNA metabarcoding. *Metabarcoding and Metagenomics*, *2*, e24262. https://doi.org/10.3897/mbmg.2.24262

Lim, N. K. M., Tay, Y. C., Srivathsan, A., Tan, J. W. T., Kwik, J. T. B., Baloğlu, B., Meier, R., & Yeo, D. C. J. (2016). Next-generation freshwater bioassessment: eDNA metabarcoding with a conserved metazoan primer reveals species-rich and reservoir-specific communities. *Royal Society Open Science*, *3*, 160635. https://doi.org/10.1098/rsos.160635

Lin, H., & Peddada, S. D. (2020). Analysis of microbial compositions: A review of normalization and differential abundance analysis. *Npj Biofilms and Microbiomes*, *6*(1), 60. https://doi.org/10.1038/s41522-020-00160-w

Lischer, H. E. L., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 474. https://doi.org/10.1186/s12859-017-1911-6

Littleford-Colquhoun, B. L., Sackett, V. I., Tulloss, C. V., & Kartzinel, T. R. (2022). Evidence-based strategies to navigate complexity in dietary DNA metabarcoding: A reply. *Molecular Ecology*, *31*(22), 5660–5665. https://doi.org/10.1111/mec.16712

Liu, J., & Zhang, H. (2021). Combining Multiple Markers in Environmental DNA Metabarcoding to Assess Deep-Sea Benthic Biodiversity. *Frontiers in Marine Science*, *8*, 684955. https://doi.org/10.3389/fmars.2021.684955

Liu, Z., Ma, A., Mathé, E., Merling, M., Ma, Q., & Liu, B. (2021). Network analyses in microbiome based on high-throughput multi-omics data. *Briefings in Bioinformatics*, *22*(2), 1639–1655. https://doi.org/10.1093/bib/bbaa005

Lopes, C. M., Sasso, T., Valentini, A., Dejean, T., Martins, M., Zamudio, K. R., & Haddad, C. F. B. (2017). eDNA metabarcoding: A promising method for anuran surveys in highly diverse tropical forests. *Molecular Ecology Resources*, *17*(5), 904–914. https://doi.org/10.1111/1755-0998.12643

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lozupone, C., & Knight, R. (2005). UniFrac: A New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., & Knight, R. (2011). UniFrac: An effective distance metric for microbial community comparison. *The ISME Journal*, *5*(2), 169–172. https://doi.org/10.1038/ismej.2010.133

Luo, M., Ji, Y., Warton, D., & Yu, D. W. (2023). Extracting abundance information from DNA -based data. *Molecular Ecology Resources*, *23*(1), 174–189. https://doi.org/10.1111/1755-0998.13703

Macher, J.-N., Vivancos, A., Piggott, J. J., Centeno, F. C., Matthaei, C. D., & Leese, F. (2018). Comparison of environmental DNA and bulk-sample metabarcoding using highly degenerate cytochrome *c* oxidase I primers. *Molecular Ecology Resources*, *18*(6), 1456–1468. https://doi.org/10.1111/1755-0998.12940

Mächler, E., Walser, J., & Altermatt, F. (2021). Decision-making and best practices for taxonomy-free environmental DNA metabarcoding in biomonitoring using Hill numbers. *Molecular Ecology*, *30*(13), 3326–3339. https://doi.org/10.1111/mec.15725

MacKenzie, D. I., & Nichols, J. D. (2004). Occupancy as a surrogate for abundance estimation. *Animal Biodiversity and Conservation*.

Magoga, G., Forni, G., Brunetti, M., Meral, A., Spada, A., De Biase, A., & Montagna, M. (2022). Curation of a reference database of COI sequences for insect identification through DNA metabarcoding: COins. *Database*, *2022*, baac055. https://doi.org/10.1093/database/baac055

Maidak, B. L., Olsen, G. J., Larsen, N., Overbeek, R., McCaughey, M. J., & Woese, C. R. (1996). The Ribosomal Database Project (RDP). *Nucleic Acids Research*, *24*(1), 82–85. https://doi.org/10.1093/nar/24.1.82

Manghi, P., Blanco-Míguez, A., Manara, S., NabiNejad, A., Cumbo, F., Beghini, F., Armanini, F., Golzato, D., Huang, K. D., Thomas, A. M., Piccinno, G., Punčochář, M., Zolfo, M., Lesker, T. R., Bredon, M., Planchais, J., Glodt, J., Valles-Colomer, M., Koren, O., … Segata, N. (2023). MetaPhlAn 4 profiling of unknown species-level genome bins improves the characterization of diet-associated microbiome changes in mice. *Cell Reports*, *42*(5), 112464. https://doi.org/10.1016/j.celrep.2023.112464

Marquardt, M., Vader, A., Stübner, E. I., Reigstad, M., & Gabrielsen, T. M. (2016). Strong Seasonality of Marine Microbial Eukaryotes in a High-Arctic Fjord (Isfjorden, in West Spitsbergen, Norway). *Applied and Environmental Microbiology*, *82*(6), 1868–1880. https://doi.org/10.1128/AEM.03208-15

Martin, C. (2011). Cutadapt removes adapter sequences for high-throughput sequencing reads. *EMBnet.Journal*, *17*, 10–12. https://doi.org/10.14806/ej.17.1.200

Martin, J. L., Santi, I., Pitta, P., John, U., & Gypens, N. (2022). Towards quantitative metabarcoding of eukaryotic plankton: An approach to improve 18S rRNA gene copy number bias. *Metabarcoding and Metagenomics*, *6*, e85794. https://doi.org/10.3897/mbmg.6.85794

Matesanz, S., Pescador, D. S., Pías, B., Sánchez, A. M., Chacón-Labella, J., Illuminati, A., Cruz, M., López-Angulo, J., Marí-Mena, N., Vizcaíno, A., & Escudero, A. (2019). Estimating belowground plant abundance with DNA metabarcoding. *Molecular Ecology Resources*, *19*(5), 1265–1277. https://doi.org/10.1111/1755-0998.13049

Matthews, S. A., Goetze, E., & Ohman, M. D. (2021). Recommendations for interpreting zooplankton metabarcoding and integrating molecular methods with morphological analyses. *ICES Journal of Marine Science*, *78*(9), 3387–3396. https://doi.org/10.1093/icesjms/fsab107

Mauffrey, F., Cordier, T., Apothéloz-Perret-Gentil, L., Cermakova, K., Merzi, T., Delefosse, M., Blanc, P., & Pawlowski, J. (2021). Benthic monitoring of oil and gas offshore platforms in the North Sea using environmental DNA metabarcoding. *Molecular Ecology*, *30*(13), 3007–3022. https://doi.org/10.1111/mec.15698

McClenaghan, B., Compson, Z. G., & Hajibabaei, M. (2020). Validating metabarcoding-based biodiversity assessments with multi-species occupancy models: A case study using coastal marine eDNA. *PLOS ONE*.

McClenaghan, B., Fahner, N., Cote, D., Chawarski, J., McCarthy, A., Rajabi, H., Singer, G., & Hajibabaei, M. (2020). Harnessing the power of eDNA metabarcoding for the detection of deep-sea fishes. *PLOS ONE*, *15*(11), e0236540. https://doi.org/10.1371/journal.pone.0236540

McInnes, J. C., Jarman, S. N., Lea, M.-A., Raymond, B., Deagle, B. E., Phillips, R. A., Catry, P., Stanworth, A., Weimerskirch, H., Kusch, A., Gras, M., Cherel, Y., Maschette, D., & Alderman, R. (2017). DNA Metabarcoding as a Marine Conservation and Management Tool: A Circumpolar Examination of Fishery Discards in the Diet of Threatened Albatrosses. *Frontiers in Marine Science*, *4*, 277. https://doi.org/10.3389/fmars.2017.00277

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (n.d.). *microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies*.

McMurdie, P. J., & Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

Medaka, O. N. T. (2018). Sequence correction provided by ONT Research. *GitHub Https://Github. Com/Nanoporetech/Medaka*.

Meglécz, E. (2022). *COInr and mkCOInr: Building and customizing a non-redundant barcoding reference database from BOLD and NCBI using a lightweight pipeline* (p. 2022.05.18.492423). bioRxiv. https://doi.org/10.1101/2022.05.18.492423

Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T. R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., Bertrand, D., Brito, J. J., Brown, C. T., Buchmann, J., Buluç, A., Chen, B., Chikhi, R., Clausen, P. T. L. C., Cristian, A., … McHardy, A. C. (2022). Critical Assessment of Metagenome Interpretation: The second round of challenges. *Nature Methods*, *19*(4), 429–440. https://doi.org/10.1038/s41592-022-01431-4

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., & Edwards, R. (2008). The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, *9*(1), 386. https://doi.org/10.1186/1471-2105-9-386

Mills, D., Fattebert, J., Hunter, L., & Slotow, R. (2019). Maximising camera trap data: Using attractants to improve detection of elusive species in multi-species surveys. *PLOS ONE*, *14*(5), e0216447. https://doi.org/10.1371/journal.pone.0216447

Miyata, K., Inoue, Y., Amano, Y., Nishioka, T., Nagaike, T., Kawaguchi, T., Morita, O., Yamane, M., & Honda, H. (2022). Comparative environmental RNA and DNA metabarcoding analysis of river algae and arthropods for ecological surveys and water quality assessment. *Scientific Reports*, *12*(1), 19828. https://doi.org/10.1038/s41598-022-23888-1

Monaghan, K. A., & Soares, A. M. V. M. (2012). Bringing new knowledge to an old problem: Building a biotic index from lotic macroinvertebrate traits. *Ecological Indicators*, *20*, 213–220. https://doi.org/10.1016/j.ecolind.2012.02.017

Morard, R., Vollmar, N. M., Greco, M., & Kucera, M. (2019). Unassigned diversity of planktonic foraminifera from environmental sequencing revealed as known but neglected species. *PLOS ONE*, *14*(3), e0213936. https://doi.org/10.1371/journal.pone.0213936

Moreno, C. E., & Halffter, G. (2000). Assessing the completeness of bat biodiversity inventories using species accumulation curves. *Journal of Applied Ecology*, *37*(1), 149–158. https://doi.org/10.1046/j.1365-2664.2000.00483.x

Muha, T. P., Rodriguez-Barreto, D., O'Rorke, R., Garcia De Leaniz, C., & Consuegra, S. (2021). Using eDNA Metabarcoding to Monitor Changes in Fish Community Composition After Barrier Removal. *Frontiers in Ecology and Evolution*, *9*, 629217. https://doi.org/10.3389/fevo.2021.629217

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H. J., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H.-H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., … Venter, J. C. (2000). A Whole-Genome Assembly of Drosophila. *Science*, *287*(5461), 2196–2204. https://doi.org/10.1126/science.287.5461.2196

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2011). MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Proceedings of the 2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, 116–124. https://doi.org/10.1145/2147805.2147818

Nasko, D. J., Koren, S., Phillippy, A. M., & Treangen, T. J. (2018). RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biology*, *19*(1), 165. https://doi.org/10.1186/s13059-018-1554-6

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), Article 5. https://doi.org/10.1038/s41587-020-00774-7

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.-M., Huntemann, M., Palaniappan, K., Ladau, J., Mukherjee, S., Reddy, T. B. K., Nielsen, T., Kirton, E., Faria, J. P., Edirisinghe, J. N., Henry, C. S., … Eloe-Fadrosh, E. A. (2021). A genomic catalog of Earth's microbiomes. *Nature Biotechnology*, *39*(4), Article 4. https://doi.org/10.1038/s41587-020-0718-6

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. (2018). Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364.

Ni, Y., Li, J., & Panagiotou, G. (2016). COMAN: A web server for comprehensive metatranscriptomics analysis. *BMC Genomics*, *17*(1), 622. https://doi.org/10.1186/s12864-016-2964-z

Nielsen, K. M., Johnsen, P. J., Bensasson, D., & Daffonchio, D. (2007). Release and persistence of extracellular DNA in the environment. *Environmental Biosafety Research*, *6*(1–2), 37–53. https://doi.org/10.1051/ebr:2007031

Nikolenko, S. I., Korobeynikov, A. I., & Alekseyev, M. A. (2013). BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics*, *14*(1), S7. https://doi.org/10.1186/1471-2164-14-S1-S7

Nissen, J. N., Johansen, J., Allesøe, R. L., Sønderby, C. K., Armenteros, J. J. A., Grønbech, C. H., Jensen, L. J., Nielsen, H. B., Petersen, T. N., Winther, O., & Rasmussen, S. (2021). Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology*, *39*(5), Article 5. https://doi.org/10.1038/s41587-020-00777-4

Nobile, M. S., Cazzaniga, P., Tangherloni, A., & Besozzi, D. (2017). Graphics processing units in bioinformatics, computational biology and systems biology. *Briefings in Bioinformatics*, *18*(5), 870–885. https://doi.org/10.1093/bib/bbw058

Numberger, D., Ganzert, L., Zoccarato, L., Mühldorfer, K., Sauer, S., Grossart, H.-P., & Greenwood, A. D. (2019). Characterization of bacterial communities in wastewater with enhanced taxonomic resolution by full-length 16S rRNA sequencing. *Scientific Reports*, *9*(1), 9673. https://doi.org/10.1038/s41598-019-46015-z

Oliveira, U., Brescovit, A. D., & Santos, A. J. (2017). Sampling effort and species richness assessment: A case study on Brazilian spiders. *Biodiversity and Conservation*, *26*(6), 1481–1493. https://doi.org/10.1007/s10531-017-1312-1

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: Fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, *16*(1), 236. https://doi.org/10.1186/s12864-015-1419-2

Ovaskainen, O., & Abrego, N. (2020). *Joint Species Distribution Modelling with Applications in R*. Cambridge University Press.

Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, *25*(5), 1032–1057. https://doi.org/10.1111/mec.13536

Pardo, I., Pata, M. P., Gómez, D., & García, M. B. (2013). A Novel Method to Handle the Effect of Uneven Sampling Effort in Biodiversity Databases. *PLoS ONE*, *8*(1), e52786. https://doi.org/10.1371/journal.pone.0052786

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research*, *50*(D1), D785–D794.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, *25*(7), 1043–1055. https://doi.org/10.1101/gr.186072.114

Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T., & Wilding, T. A. (2014). Environmental monitoring through protist next-generation sequencing metabarcoding: Assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources*, *14*(6), 1129–1140. https://doi.org/10.1111/1755-0998.12261

Pawlowski, J., Esling, P., Lejzerowicz, F., Cordier, T., Visco, J., Martins, C., Kvalvik, A., Staven, K., & Cedhagen, T. (2016). Benthic monitoring of salmon farms in Norway using foraminiferal metabarcoding. *Aquaculture Environment Interactions*, *8*, 371–386. https://doi.org/10.3354/aei00182

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M. J., Filipe, A. F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Iwan Jones, J., Sagova-Mareckova, M., Moritz, C., … Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of The Total Environment*, *637–638*, 1295–1310. https://doi.org/10.1016/j.scitotenv.2018.05.002

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2011). Meta-IDBA: A de Novo assembler for metagenomic data. *Bioinformatics*, *27*(13), i94–i101. https://doi.org/10.1093/bioinformatics/btr216

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, *28*(11), 1420–1428. https://doi.org/10.1093/bioinformatics/bts174

Pierella Karlusich, J. J., Pelletier, E., Zinger, L., Lombard, F., Zingone, A., Colin, S., Gasol, J. M., Dorrell, R. G., Henry, N., Scalco, E., Acinas, S. G., Wincker, P., De Vargas, C., & Bowler, C. (2023). A robust approach to estimate relative phytoplankton cell abundances from metagenomes. *Molecular Ecology Resources*, *23*(1), 16–40. https://doi.org/10.1111/1755-0998.13592

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2018). The choice of universal primers and the characteristics of the species mixture determines when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(2). https://doi.org/10.1111/mec.14776

Pochon, X., Wood, S. A., Keeley, N. B., Lejzerowicz, F., Esling, P., Drew, J., & Pawlowski, J. (2015). Accurate assessment of the impact of salmon farming on benthic sediment enrichment using foraminiferal metabarcoding. *Marine Pollution Bulletin*, *100*(1), 370–382. https://doi.org/10.1016/j.marpolbul.2015.08.022

Poncheewin, W., Hermes, G. D., Van Dam, J. C., Koehorst, J. J., Smidt, H., & Schaap, P. J. (2020). NG-Tax 2.0: A semantic framework for high-throughput amplicon analysis. *Frontiers in Genetics*, *10*, 1366.

Pont, D., Meulenbroek, P., Bammer, V., Dejean, T., Erős, T., Jean, P., Lenhardt, M., Nagel, C., Pekarik, L., Schabuss, M., Stoeckle, B. C., Stoica, E., Zornig, H., Weigand, A., & Valentini, A. (2023). Quantitative monitoring of diverse fish communities on a large scale combining EDNA metabarcoding and QPCR. *Molecular Ecology Resources*, *23*(2), 396–409. https://doi.org/10.1111/1755-0998.13715

Pont, D., Rocle, M., Valentini, A., Civade, R., Jean, P., Maire, A., Roset, N., Schabuss, M., Zornig, H., & Dejean, T. (2018). Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. *Scientific Reports*, *8*(1), 1–13. https://doi.org/10.1038/s41598-018-28424-8

Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PloS One*, *13*(9), e0200177.

Porter, T. M., & Hajibabaei, M. (2020). Putting COI metabarcoding in context: The utility of exact sequence variants (ESVs) in biodiversity analysis. *Frontiers in Ecology and Evolution*, *8*, 248.

Porter, T. M., & Hajibabaei, M. (2021). Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics*, *22*(1), 256. https://doi.org/10.1186/s12859-021-04180-x

Porter, T. M., Morris, D. M., Basiliko, N., Hajibabaei, M., Doucet, D., Bowman, S., Emilson, E. J. S., Emilson, C. E., Chartrand, D., Wainio-Keizer, K., Séguin, A., & Venier, L. (2019). Variations in terrestrial arthropod DNA metabarcoding methods recovers robust beta diversity but variable richness and site indicators. *Scientific Reports*, *9*(1), 18218. https://doi.org/10.1038/s41598-019-54532-0

Processor Specifications | AMD. (n.d.). Retrieved January 28, 2024, from https://www.amd.com/en/products/specifications/processors

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(suppl_1), D61–D65.

Pukk, L., Kanefsky, J., Heathman, A. L., Weise, E. M., Nathan, L. R., Herbst, S. J., Sard, N. M., Scribner, K. T., & Robinson, J. D. (2021). eDNA metabarcoding in lakes to quantify influences of landscape features and human activity on aquatic invasive species prevalence and fish community diversity. *Diversity and Distributions*, *27*(10), 2016–2031. https://doi.org/10.1111/ddi.13370

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), D590–D596. https://doi.org/10.1093/nar/gks1219

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., & Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, *35*(9), Article 9. https://doi.org/10.1038/nbt.3935

Ratcliffe, F. C., Uren Webster, T. M., Rodriguez-Barreto, D., O'Rorke, R., Garcia De Leaniz, C., & Consuegra, S. (2021). Quantitative assessment of fish larvae community composition in spawning areas using metabarcoding of bulk samples. *Ecological Applications*, *31*(3). https://doi.org/10.1002/eap.2284

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (http://www. Barcodinglife. Org). *Molecular Ecology Notes*, *7*(3), 355–364.

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., & Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, *8*(1), 64–77. https://doi.org/10.1007/s40484-019-0187-4

Ritter, C. D., Forster, D., Azevedo, J. A. R., Antonelli, A., Nilsson, R. H., Trujillo, M. E., & Dunthorn, M. (2021). Assessing Biotic and Abiotic Interactions of Microorganisms in Amazonia through Co-Occurrence Networks and DNA Metabarcoding. *Microbial Ecology*, *82*(3), 746–760. https://doi.org/10.1007/s00248-021-01719-6

Roberts, D. W. (2020). Comparison of distance-based and model-based ordinations. *Ecology*, *101*(1). https://doi.org/10.1002/ecy.2908

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. (2021). RESCRIPt: Reproducible sequence taxonomy reference database management. *PLoS Computational Biology*, *17*(11), e1009581.

Robinson, C. V., Porter, T. M., McGee, K. M., McCusker, M., Wright, M. T. G., & Hajibabaei, M. (2022). Multi-marker DNA metabarcoding detects suites of environmental gradients from an urban harbour. *Scientific Reports*, *12*(1), Article 1. https://doi.org/10.1038/s41598-022-13262-6

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Roswell, M., Dushoff, J., & Winfree, R. (2021). A conceptual guide to measuring species diversity. *Oikos*, *130*(3), 321–338. https://doi.org/10.1111/oik.07202

Rourke, M. L., Fowler, A. M., Hughes, J. M., Broadhurst, M. K., DiBattista, J. D., Fielder, S., Wilkes Walburn, J., & Furlan, E. M. (2022). Environmental DNA (eDNA) as a tool for assessing fish biomass: A review of approaches and future considerations for resource surveys. *Environmental DNA*, *4*(1), 9–33. https://doi.org/10.1002/edn3.185

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*, e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Ruscheweyh, H.-J., Milanese, A., Paoli, L., Sintsova, A., Mende, D. R., Zeller, G., & Sunagawa, S. (2021). mOTUs: Profiling Taxonomic Composition, Transcriptional Activity and Strain Populations of Microbial Communities. *Current Protocols*, *1*(8), e218. https://doi.org/10.1002/cpz1.218

Saary, P., Mitchell, A. L., & Finn, R. D. (2020). Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biology*, *21*(1), 244. https://doi.org/10.1186/s13059-020-02155-4

Sanchez, L., Boulanger, E., Arnal, V., Boissery, P., Dalongeville, A., Dejean, T., Deter, J., Guellati, N., Holon, F., Juhel, J.-B., Lenfant, P., Leprieur, F., Valentini, A., Manel, S., & Mouillot, D. (2022). Ecological indicators based on quantitative eDNA metabarcoding: The case of marine reserves. *Ecological Indicators*, *140*, 108966. https://doi.org/10.1016/j.ecolind.2022.108966

Santoferrara, L. F. (2019). Current practice in plankton metabarcoding: Optimization and error management. *Journal of Plankton Research*, *41*(5), 571–582. https://doi.org/10.1093/plankt/fbz041

Sard, N. M., Herbst, S. J., Nathan, L., Uhrig, G., Kanefsky, J., Robinson, J. D., & Scribner, K. T. (2019). Comparison of fish detections, community diversity, and relative abundance using environmental DNA metabarcoding and traditional gears. *Environmental DNA*, *1*(4), 368–384. https://doi.org/10.1002/edn3.38

Schenekar, T., Schletterer, M., Lecaudey, L. A., & Weiss, S. J. (2020). Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. *River Research and Applications*, rra.3610. https://doi.org/10.1002/rra.3610

Schenk, J., Geisen, S., Kleinboelting, N., & Traunspurger, W. (2019). Metabarcoding data allow for reliable biomass estimates in the most abundant animals on earth. *Metabarcoding and Metagenomics*, *3*, e46704. https://doi.org/10.3897/mbmg.3.46704

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., & Robinson, C. J. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541.

Schmidt, B. R., Kéry, M., Ursenbacher, S., Hyman, O. J., & Collins, J. P. (2013). Site occupancy models in the analysis of environmental DNA presence/absence surveys: A case study of an emerging amphibian pathogen. *Methods in Ecology and Evolution*, *4*(7), 646–653. https://doi.org/10.1111/2041-210X.12052

Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., & Goesmann, A. (2021). Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics*, *7*(11), 000685. https://doi.org/10.1099/mgen.0.000685

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., … McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation—A benchmark of metagenomics software. *Nature Methods*, *14*(11), 1063–1071. https://doi.org/10.1038/nmeth.4458

Shelton, A. O., Gold, Z. J., Jensen, A. J., D'Agnese, E., Andruszkiewicz, E., & Kelly, P. (2022). Toward quantitative metabarcoding. *Ecology*, *104*(2). https://doi.org/10.1002/ecy.3906

Shirazi, S., Meyer, R. S., & Shapiro, B. (2021). Revisiting the effect of PCR replication and sequencing depth on biodiversity metrics in environmental DNA metabarcoding. *Ecology and Evolution*, *11*(22), 15766–15779. https://doi.org/10.1002/ece3.8239

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., & Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: A case study of eDNA metabarcoding seawater. *Scientific Reports*, *9*, 5991. https://doi.org/10.1038/s41598-019-42455-9

Singer, G. A. C., Shekarriz, S., McCarthy, A., Fahner, N., & Hajibabaei, M. (2020). *The utility of a metagenomics approach for marine biomonitoring* (p. 2020.03.16.993667). bioRxiv. https://doi.org/10.1101/2020.03.16.993667

Skelton, J., Cauvin, A., & Hunter, M. E. (2022). Environmental DNA metabarcoding read numbers and their variability predict species abundance, but weakly in non-dominant species. *Environmental DNA*, edn3.355. https://doi.org/10.1002/edn3.355

Skidmore, A. K., Siegenthaler, A., Wang, T., Darvishzadeh, R., Zhu, X., Chariton, A., & Arjen De Groot, G. (2022). Mapping the relative abundance of soil microbiome biodiversity from eDNA and remote sensing. *Science of Remote Sensing*, *6*, 100065. https://doi.org/10.1016/j.srs.2022.100065

Smets, W., Leff, J. W., Bradford, M. A., McCulley, R. L., Lebeer, S., & Fierer, N. (2016). A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology and Biochemistry*, *96*, 145–151.

Somervuo, P., Koskela, S., Pennanen, J., Henrik Nilsson, R., & Ovaskainen, O. (2016). Unbiased probabilistic taxonomic classification for DNA barcoding. *Bioinformatics*, *32*(19), 2920–2927. https://doi.org/10.1093/bioinformatics/btw346

Staehr, P. A. U., Dahl, K., Buur, H., Göke, C., Sapkota, R., Winding, A., Panova, M., Obst, M., & Sundberg, P. (2022). Environmental DNA Monitoring of Biodiversity Hotspots in Danish Marine Waters. *Frontiers in Marine Science*, *8*, 800474. https://doi.org/10.3389/fmars.2021.800474

Stefanni, S., Stanković, D., Borme, D., De Olazabal, A., Juretić, T., Pallavicini, A., & Tirelli, V. (2018). Multi-marker metabarcoding approach to study mesozooplankton at basin scale. *Scientific Reports*, *8*(1), 12085. https://doi.org/10.1038/s41598-018-30157-7

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., & DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, *178*(3), 591–599. https://doi.org/10.1128/jb.178.3.591-599.1996

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), Article 11. https://doi.org/10.1038/nbt.3988

Stier, A. C., Bolker, B. M., & Osenberg, C. W. (2016). Using rarefaction to isolate the effects of patch size and sampling effort on beta diversity. *Ecosphere*, *7*(12). https://doi.org/10.1002/ecs2.1612

Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., & Pawlowski, J. (2018). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecological Indicators*, *85*, 153–164. https://doi.org/10.1016/j.ecolind.2017.10.041

Stoeck, T., Pan, H., Dully, V., Forster, D., & Jung, T. (2018). Towards an eDNA metabarcode-based performance indicator for full-scale municipal wastewater treatment plants. *Water Research*, *144*, 322–331. https://doi.org/10.1016/j.watres.2018.07.051

Stolar, J., & Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, *21*(5), 595–608. https://doi.org/10.1111/ddi.12279

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M. B., Wincker, P., & de Vargas, C. (2020). Tara Oceans: Towards global ocean ecosystems biology. *Nature Reviews Microbiology*, *18*(8), Article 8. https://doi.org/10.1038/s41579-020-0364-5

Suter, L., Polanowski, A. M., Clarke, L. J., Kitchener, J. A., & Deagle, B. E. (2021). Capturing open ocean biodiversity: Comparing environmental DNA metabarcoding to the continuous plankton recorder. *Molecular Ecology*, *30*(13), 3140–3157. https://doi.org/10.1111/mec.15587

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018a). DNA metabarcoding data analysis. In P. Taberlet, A. Bonin, L. Zinger, & E. Coissac (Eds.), *Environmental DNA: For Biodiversity Research and Monitoring* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780198767220.003.0008

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018b). Reference databases. In P. Taberlet, A. Bonin, L. Zinger, & E. Coissac (Eds.), *Environmental DNA: For Biodiversity Research and Monitoring* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780198767220.003.0003

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018c). Terrestrial ecosystems. In P. Taberlet, A. Bonin, L. Zinger, & E. Coissac (Eds.), *Environmental DNA: For Biodiversity Research and Monitoring* (p. 0). Oxford University Press. https://doi.org/10.1093/oso/9780198767220.003.0014

Takeuchi, A., Iijima, T., Kakuzen, W., Watanabe, S., Yamada, Y., Okamura, A., Horie, N., Mikawa, N., Miller, M. J., Kojima, T., & Tsukamoto, K. (2019). Release of eDNA by different life history stages and during spawning activities of laboratory-reared Japanese eels for interpretation of oceanic survey data. *Scientific Reports*, *9*(1), 6074. https://doi.org/10.1038/s41598-019-42641-9

Tamames, J., & Puente-Sánchez, F. (2019). SqueezeMeta, A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline. *Frontiers in Microbiology*, *9*, 3349. https://doi.org/10.3389/fmicb.2018.03349

Tapolczai, K., Selmeczy, G. B., Szabó, B., B-Béres, V., Keck, F., Bouchez, A., Rimet, F., & Padisák, J. (2021). The potential of exact sequence variants (ESVs) to interpret and assess the impact of agricultural pressure on stream diatom assemblages revealed by DNA metabarcoding. *Ecological Indicators*, *122*, 107322. https://doi.org/10.1016/j.ecolind.2020.107322

Tarazona, S., Furió-Tarí, P., Turrà, D., Pietro, A. D., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, *43*(21), e140. https://doi.org/10.1093/nar/gkv711

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R. H., Kennedy, P. G., Yang, T., Anslan, S., & Mikryukov, V. (2022). Best practices in metabarcoding of fungi: From experimental design to results. *Molecular Ecology*, *31*(10), 2769–2795. https://doi.org/10.1111/mec.16460

Telenius, A. (2011). Biodiversity information goes public: GBIF at your service. *Nordic Journal of Botany*, *29*(3), 378–381.

Ter Braak, C. J. F., & Šmilauer, P. (2015). Topics in constrained and unconstrained ordination. *Plant Ecology*, *216*(5), 683–696. https://doi.org/10.1007/s11258-014-0356-5

Thomas, A. C., Deagle, B. E., Eveson, J. P., Harsch, C. H., & Trites, A. W. (2016). Quantitative DNA metabarcoding: Improved estimates of species proportional biomass using correction factors derived from control material. *Molecular Ecology Resources*, *16*(3), 714–726. https://doi.org/10.1111/1755-0998.12490

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., Jonge, M. M. J., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the R -package H MSC. *Methods in Ecology and Evolution*, *11*(3), 442–447. https://doi.org/10.1111/2041-210X.13345

Tobler, M. W., Kéry, M., Hui, F. K. C., Guillera-Arroita, G., Knaus, P., & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, *100*(8). https://doi.org/10.1002/ecy.2754

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., & Pop, M. (2011). Next Generation Sequence Assembly with AMOS. *Current Protocols in Bioinformatics*, *33*(1), 11.8.1-11.8.18. https://doi.org/10.1002/0471250953.bi1108s33

Tsuji, S., Inui, R., Nakao, R., Miyazono, S., Saito, M., Kono, T., & Akamatsu, Y. (2022). Quantitative environmental DNA metabarcoding shows high potential as a novel approach to quantitatively assess fish community. *Scientific Reports*, *12*(1), 21524. https://doi.org/10.1038/s41598-022-25274-3

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020). Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA*, *2*(1), 42–52. https://doi.org/10.1002/edn3.44

Tuomisto, H. (2010a). A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia*, *164*(4), 853–860. https://doi.org/10.1007/s00442-010-1812-0

Tuomisto, H. (2010b). A diversity of beta diversities: Straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*, *33*(1), 2–22. https://doi.org/10.1111/j.1600-0587.2009.05880.x

Van Bleijswijk, J. D. L., Engelmann, J. C., Klunder, L., Witte, H. J., Witte, J. Ij., & Van Der Veer, H. W. (2020). Analysis of a coastal North Sea fish community: Comparison of aquatic environmental DNA concentrations to fish catches. *Environmental DNA*, *2*(4), 429–445. https://doi.org/10.1002/edn3.67

Van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The third revolution in sequencing technology. *Trends in Genetics*, *34*(9), 666–681.

Veldsman, W. P., Campli, G., Dind, S., Rech de Laval, V., Drage, H., Waterhouse, R. M., & Robinson-Rechavi, M. (2022). Taxonbridge: An R package to create custom taxonomies based on the NCBI and GBIF taxonomies. *bioRxiv*, 2022–05.

Visco, J. A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L., & Pawlowski, J. (2015). Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data. *Environmental Science & Technology*, *49*(13), 7597–7605. https://doi.org/10.1021/es506158m

Wangensteen, O. S., Palacín, C., Guardiola, M., & Turon, X. (2018). DNA metabarcoding of littoral hard-bottom communities: High diversity and database gaps revealed by two molecular markers. *PeerJ*, *6*, e4705. https://doi.org/10.7717/peerj.4705

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, *30*(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J., & Dunstan, P. K. (2015). Model-based thinking for community ecology. *Plant Ecology*, *216*(5), 669–682. https://doi.org/10.1007/s11258-014-0366-3

Warton, D. I., Wright, S. T., & Wang, Y. (n.d.). Distance-based multivariate analyses confound location and dispersion effects: *Mean-variance confounding in multivariate analysis*. *Methods in Ecology and Evolution*, *3*(1), 89–101. https://doi.org/10.1111/j.2041-210X.2011.00127.x

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., & Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, *5*(1), 27. https://doi.org/10.1186/s40168-017-0237-y

Wemheuer, F., Taylor, J. A., Daniel, R., Johnston, E., Meinicke, P., Thomas, T., & Wemheuer, B. (2020). Tax4Fun2: Prediction of habitat-specific functional profiles and functional redundancy based on 16S rRNA gene sequences. *Environmental Microbiome*, *15*(1), 11. https://doi.org/10.1186/s40793-020-00358-7

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., Travers, M. J., Newman, S. J., & Bunce, M. (2020). eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. *Molecular Ecology*, *29*(6), 1069–1086. https://doi.org/10.1111/mec.15382

West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C., & Banfield, J. F. (2018). Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Research*, *28*(4), 569–580. https://doi.org/10.1101/gr.228429.117

Westreich, S. T., Treiber, M. L., Mills, D. A., Korf, I., & Lemay, D. G. (2018). SAMSA2: A standalone metatranscriptome analysis pipeline. *BMC Bioinformatics*, *19*(1), 175. https://doi.org/10.1186/s12859-018-2189-z

Wick, R. (2018). *Porechop: Adapter trimmer for Oxford Nanopore reads*.

Wick, R. R., Judd, L. M., & Holt, K. E. (2019). Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology*, *20*, 1–10.

Wilding, T. A., Stoeck, T., Morrissey, B. J., Carvalho, S. F., & Coulson, M. W. (2023). Maximising signal-to-noise ratios in environmental DNA-based monitoring. *Science of The Total Environment*, *858*, 159735. https://doi.org/10.1016/j.scitotenv.2022.159735

Willoughby, J. R., Wijayawardena, B. K., Sundaram, M., Swihart, R. K., & DeWoody, J. A. (2016). The importance of including imperfect detection models in eDNA experimental design. *Molecular Ecology Resources*, *16*(4), 837–844. https://doi.org/10.1111/1755-0998.12531

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wood, S. A., Biessy, L., Latchford, J. L., Zaiko, A., Von Ammon, U., Audrezet, F., Cristescu, M. E., & Pochon, X. (2020). Release and degradation of environmental DNA and RNA in a marine system. *Science of The Total Environment*, *704*, 135314. https://doi.org/10.1016/j.scitotenv.2019.135314

Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607. https://doi.org/10.1093/bioinformatics/btv638

Xia, Y., & Sun, J. (2017). Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases*, *4*(3), 138–148. https://doi.org/10.1016/j.gendis.2017.06.001

Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T., & Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, *7*(1). https://doi.org/10.1038/srep40368

Yang, C., Wang, X., Miller, J. A., De Blécourt, M., Ji, Y., Yang, C., Harrison, R. D., & Yu, D. W. (2014). Using metabarcoding to ask if easily collected soil and leaf-litter samples can be used as a general biodiversity indicator. *Ecological Indicators*, *46*, 379–389. https://doi.org/10.1016/j.ecolind.2014.06.028

Yang, J., Zhang, X., Xie, Y., Song, C., Zhang, Y., Yu, H., & Burton, G. A. (2017). Zooplankton Community Profiling in a Eutrophic Freshwater Ecosystem-Lake Tai Basin by DNA Metabarcoding. *Scientific Reports*, *7*(1), 1773. https://doi.org/10.1038/s41598-017-01808-y

Yates, M. C., Wilcox, T. M., Stoeckle, M. Y., & Heath, D. D. (2022). Interspecific allometric scaling in eDNA production among northwestern Atlantic bony fishes reflects physiological allometric scaling. *Environmental DNA*, edn3.381. https://doi.org/10.1002/edn3.381

Ye, S. H., Siddle, K. J., Park, D. J., & Sabeti, P. C. (2019). Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell*, *178*(4), 779–794. https://doi.org/10.1016/j.cell.2019.07.010

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P.,

Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, *29*(5), 415–420. https://doi.org/10.1038/nbt.1823

Zemb, O., Achard, C. S., Hamelin, J., De Almeida, M., Gabinaud, B., Cauquil, L., Verschuren, L. M. G., & Godon, J. (2020). Absolute quantitation of microbes using 16S rRNA gene metabarcoding: A rapid normalization of relative abundances by quantitative PCR targeting a 16S rRNA gene spike-in standard. *MicrobiologyOpen*, *9*(3), e977. https://doi.org/10.1002/mbo3.977

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829. https://doi.org/10.1101/gr.074492.107

Zhang, Y., Thompson, K. N., Branck, T., Yan Yan, null, Nguyen, L. H., Franzosa, E. A., & Huttenhower, C. (2021). Metatranscriptomics for the Human Microbiome and Microbial Community Functional Profiling. *Annual Review of Biomedical Data Science*, *4*, 279–311. https://doi.org/10.1146/annurev-biodatasci-031121-103035

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *7*(1–2), 203–214. https://doi.org/10.1089/10665270050081478

Zhou, S., Fan, C., Xia, H., Zhang, J., Yang, W., Ji, D., Wang, L., Chen, L., & Liu, N. (2022). Combined Use of eDNA Metabarcoding and Bottom Trawling for the Assessment of Fish Biodiversity in the Zhoushan Sea. *Frontiers in Marine Science*, *8*, 809703. https://doi.org/10.3389/fmars.2021.809703

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding—Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. https://doi.org/10.1111/mec.15060

8 Appendix A: Occupancy Models

eDNA surveys typically involve collection of multiple samples per site location (e.g., environmental replicates) and laboratory analysis includes subsampling of the eDNA extract from each individual sample (e.g., qPCR technical replicates). Therefore, eDNA surveys typically include three nested levels of sampling (Figure 14):

1. Locations or sites (primary sample units) within a study area,
2. Environmental samples (secondary sample units) collected form each location, and
3. Subsamples (replicate observations – qPCR) taken within each environmental sample

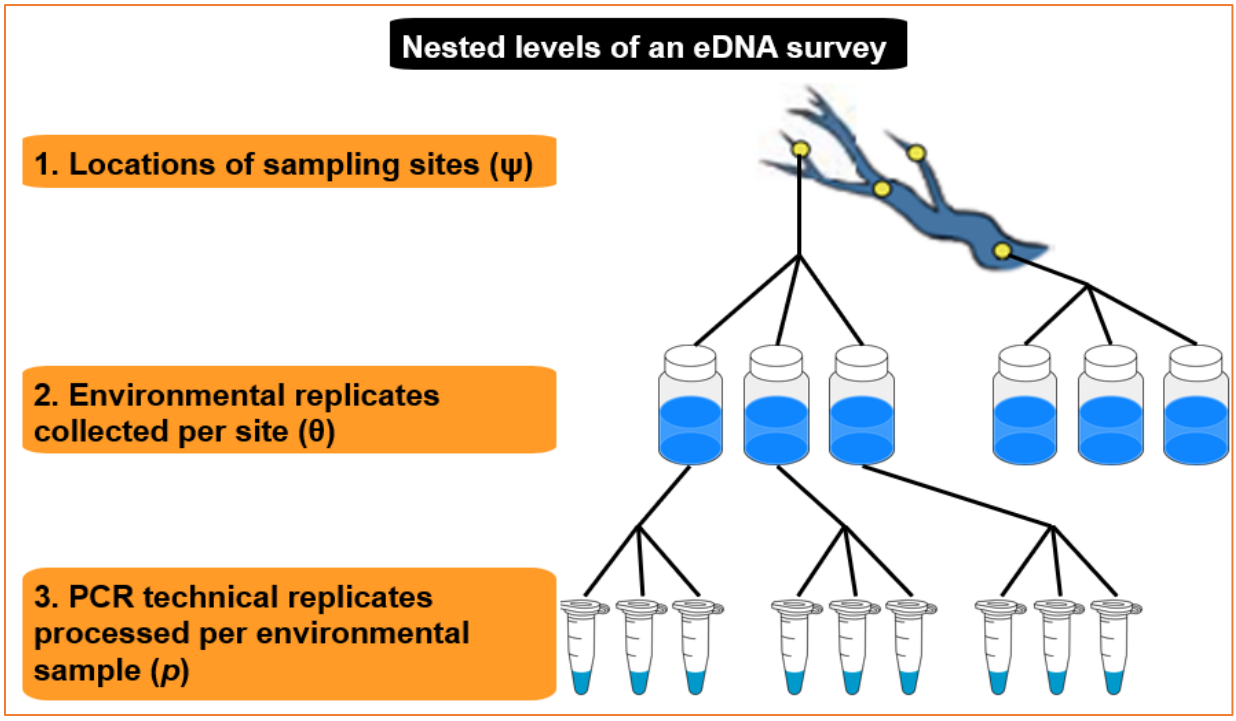


Figure 14: ***Diagram depicting the nested levels typically implemented in an eDNA survey. The first level indicates sampling locations or sites where eDNA is collected. The second level indicates multiple environmental replicates collected per site. The third level indicates multiple PCR replicates processed per environmental sample collected.***

Therefore, a multiscale occupancy model can be implemented to estimate (Figure 14):

1. the probability of target species occurrence at the location (ψ, *psi*),
2. the conditional probability of target eDNA occurrence in an environmental sample given that the target species is present at that location (ϑ, *theta*), and
3. the conditional probability of positive detection in a qPCR replicate given that the target eDNA is present in the environmental sample (p).

The simplest model for this multiscale occupancy ($\psi(.)$, $\vartheta(.)$, $p(.)$) estimates the mean probability the species is present across any of the sampling locations (ψ), the mean probability target eDNA was

collected in a sample if the target species was present (ϑ), and the mean probability eDNA was detected in a qPCR replicate if the target eDNA was collected (p).

- Using the equation $1-(1-\vartheta)^n \geq 0.95$, where ϑ is the probability of eDNA occurrence and n is the number of water samples, the number of water samples required to surpass 95% probability of successful eDNA collection can be calculated,
- Using the equation $1-(1-p)^n \geq 0.95$, where p is the probability of qPCR detection and n is the number of qPCR replicates, the number of qPCR replicates required to surpass 95% probability of detection within a water sample can be calculated.

Both the probability of eDNA collection (ϑ) and the probability of qPCR detection (p) will influence the cumulative probability of eDNA detection (Figure 15). For example, a qPCR cannot detect an eDNA molecule that was never successfully collected, no matter how many qPCR replicates are performed. Collecting more eDNA samples can improve the probability of eDNA collection and performing more qPCR analysis can increase the probability of eDNA detection within a qPCR. A pilot study performed at sites of known species occurrence can help to evaluate what level of effort (both environmental and qPCR replicates) is required to confidently conclude presence or probable absence from an eDNA survey.

A species may have a low probability of eDNA occurrence in a water sample (ϑ) if it is rare, displays a low eDNA shed rate, occurs in a habitat with large water volume (causing eDNA dilution effects), or if samples aren't collected in close proximity to individuals (e.g., surface water samples collected targeting a benthic organism). Furthermore, differences in environmental covariates across sampling sites are likely to influence the probability of eDNA occurrence in a water sample (ϑ), for example turbidity (reducing the amount volume filtered), water discharge/flow, river width or depth, and many more.

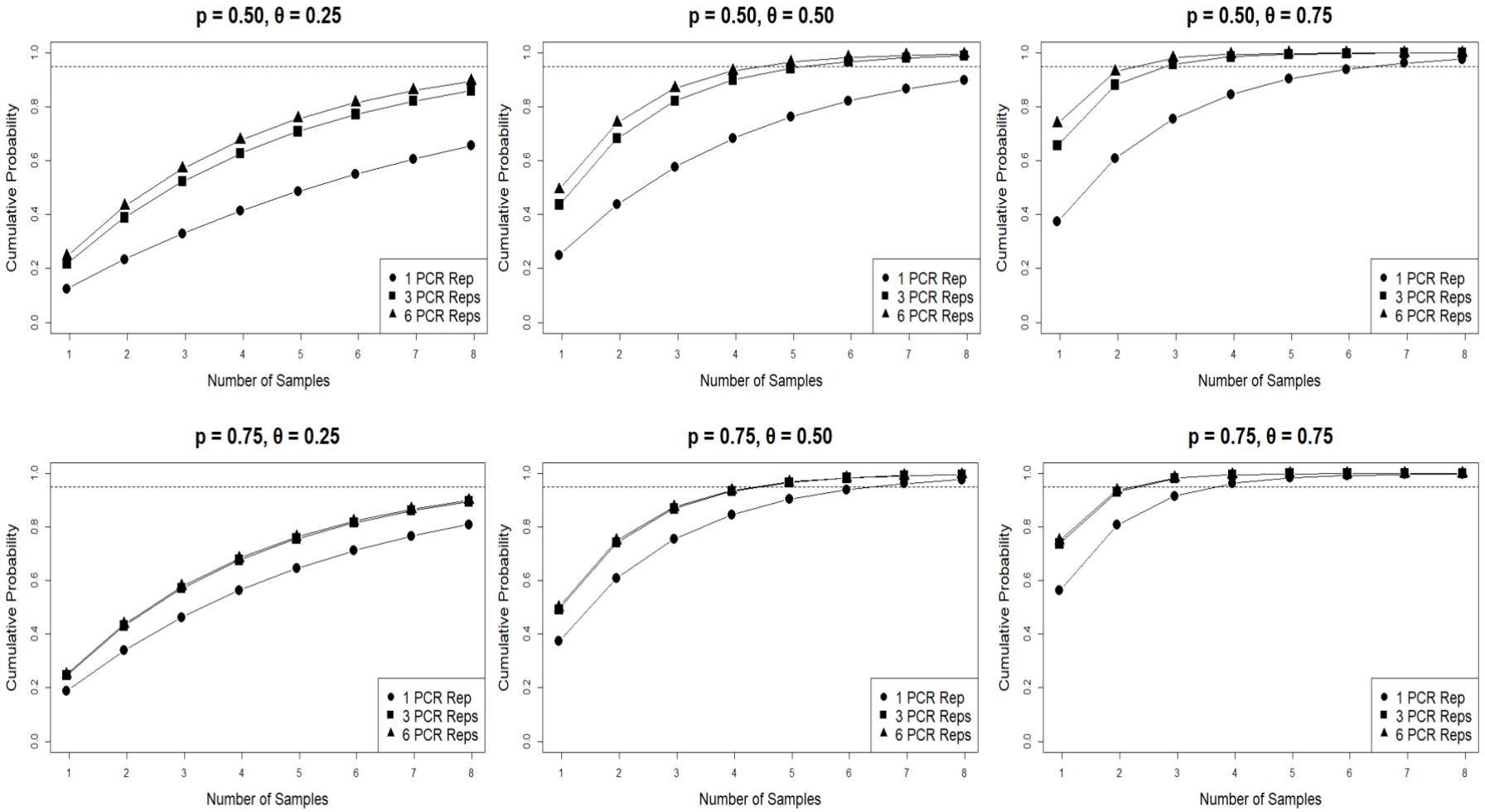


Figure 15: ***Changes to the cumulative probability of eDNA detection based on the number of environmental samples collected. Each plot displays differences across an increasing number of PCR replicates (1, 3, or 6). The top three plots display probabilities based on a PCR detection of p = 0.50 and the bottom three display probabilities based on PCR detection of p = 0.75. Plots from left to right indicate cumulative probabilities based on increasing probability of eDNA collection (ϑ = 0.25, 0.50, or 0.75). Dashed line indicates a 95% probability of eDNA detection.***

Through occupancy models, the ϑ and p probabilities can be estimated and the sampling design can be evaluated to ensure sufficient sampling to reach a target probability of species detection. This can provide confidence to an eDNA study program. Let's take an example where the probability of eDNA collection within a sample (ϑ) is 0.50 and the probability of eDNA detection with a PCR replicate (p) is 0.75 (Figure C.2). With these parameters, we can see how changing the number of samples collected (the x-axis of Figure C.2) or changing the number of PCR replicates (the three curves in Figure C.2) impact the cumulative probability of detection. In this scenario, it is required to collect five eDNA samples and analyze at least three PCR replicates to achieve a >95% probability of detection (dashed line in Figure 15).

Several software packages have been developed for occupancy modeling analysis of eDNA data:

- eDNAoccupancy R package (Dorazio & Erickson 2017) – most commonly used,
- msocc R package (Stratton et al. 2020),
- https://seak.shinyapps.io/eDNA/ (Griffin et al. 2020) – recently developed to incorporate probability of false positives.

These programs typically perform a Bayesian multiscale occupancy model to estimate posterior summaries of occurrence and detection probabilities. Furthermore, a user can evaluate how occurrence and detection probabilities are impacted from environmental and sampling covariates. This allows users to evaluate if certain environmental or sampling covariates are impacting the probability of species presence (ψ) and/or the probability of successful detection (θ and p). This information can help to both understand what factors impact the distribution of a species across sites/locations, and what factors impact the successful collection and detection of eDNA for the target species. Furthermore, analysis of occupancy models can provide estimates of probability of collection within a water sample (θ) and probability of detection within a qPCR replicate (p), to inform if appropriate sampling effort was undertaken. It is important to collect the necessary environmental metadata during sampling that an end user wants to evaluate in an occupancy model.