NEMO TOOLS

Next-generation monitoring & mapping tools to assess marine ecosystems & biodiversity

Deliverable D2.2

Protocol for the taxonomic and functional identification of species of the different biocommunities examined

Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN

Funded by the European Union
NextGenerationEU

This project is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0, funded by the European Union – NextGenerationEU (Implementation body: HFRI).

DOCUMENT INFORMATION AND VERSION CONTROL

Project Acronym	NEMO-Tools
Project Title	Next-generation monitoring and mapping tools to assess marine ecosystems and biodiversity
Project Number	016035
Work Package	WP2
Related Task(s)	T2.2
Deliverable Number	D2.2
Deliverable Name	Protocol for the taxonomic and functional identification of species of the different biocommunities examined
Due Date	14 February 2024
Date Delivered	
Dissemination Level	Public — fully open (automatically posted online on the Project Results platforms)

VERSION CONTROL

Revision-N°	Date	Description	Prepared By	Reviewed By
	10/01/2025	1st Draft	S. Genitsaris	C. Gubili
		Final Draft		

Executive Summary

This Deliverable 2.2 – "Protocol of the taxonomic and functional identification of species of the different biocommunities examined" provides a step-by-step standard operating procedure to analyze raw sequencing amplicon reads towards taxonomic synthesis and functional prediction of bacterial, microeukaryotic and fish communities using bioinformatics. This document describes the tools and downstream analysis pipelines that are recommended for the produced sequencing datasets based on the research team's large experience and on state-of-the-art techniques and on similar high-throughput data, commonly used in literature. The described framework will be implemented in the following tasks of WP2 and are crucial to achieve the overarching ecological endpoints of NEMO-Tools.

TABLE OF CONTENTS

DOCUMENT INFORMATION AND VERSION CONTROL 3
VERSION CONTROL 3
Executive Summary 4
TABLE OF CONTENTS 5
CONTRIBUTORS 6
1. Introduction 7
1. eDNA extraction from filters 9
2. Microbial Communities 9
3. Fish 12
4. Challenges of eDNA 15
5. Perspectives 15
6. References 17

CONTRIBUTORS

TABLE 1 NAMES AND ROLES OF CONTRIBUTORS TO THIS DELIVERABLE.

Name	Affiliation	WP Lead	Task Lead
Chrysoula Gubili	Hellenic Agricultural Organization – DIMITRA - Fisheries Research Institute	2	
Savvas Genitsaris	National and Kapodistrian University of Athens		2.2
Panagiota Xanthopoulou	Hellenic Agricultural Organization – DIMITRA - Fisheries Research Institute		
Antonios Mazaris	Aristotle University of Thessaloniki		

1. Introduction

Understanding the complexity of marine ecosystems is crucial for conservation and sustainable management. Marine environments harbour diverse communities, from the microscopic to the macroscopic world, that contribute to ecosystem functioning and health, including nutrient cycling, primary production, and habitat formation (Thomsen & Willerslev, 2015). The advent of high-throughput sequencing (HTS) technologies has revolutionized biodiversity studies, enabling researchers to assess the taxonomic and functional diversity of marine organisms efficiently. By analysing environmental DNA (eDNA) through metabarcoding, it is now possible to detect multiple taxa from a single environmental sample, providing a snapshot of the biodiversity within a given ecosystem (Stat *et al.*, 2017).

However, the rapid advancements in HTS technologies have led to an unprecedented explosion of sequencing data. Platforms such as Illumina have enabled the generation of terabytes of data within days, with applications spanning metabarcoding, genomics, and metagenomics. For instance, the global sequencing data output is estimated to exceed exabytes annually, driven by large-scale projects such as marine biomonitoring initiatives (Goodwin *et al.*, 2016; Reuter *et al.*, 2015). This surge in data has created an urgent need for powerful and scalable bioinformatics tools to process, analyse, and interpret vast datasets. Raw sequencing reads contain errors, artifacts, and contaminants that require comprehensive workflows for quality control, assembly, annotation, and functional interpretation, as well as the integration of multi-omics datasets to gain meaningful insights. Tools like Mothur, QIIME2, and DADA2 have been developed to meet these demands in metabarcoding outputs, but continued innovation is essential to keep pace with the ever-increasing complexity and volume of NGS data (Schloss *et al.*, 2009; Bolyen *et al.*, 2019; Callahan *et al.*, 2016). Bioinformatic cleaning is therefore a critical step to ensure data quality and reliability.

In the literature, numerous software have been introduced to mainly perform the following steps that are key for the assembly of rigorous biodiversity datasets:

1. **Quality Control**: Quality assessment tools like FastQC (Andrews, 2010) provide insights into sequence quality, GC content, and adapter contamination. Based on this, low-quality sequences can be identified.
2. **Trimming and Filtering**: Tools like Trimmomatic (Bolger *et al.*, 2014) or Cutadapt (Martin, 2011) are widely used to remove sequencing adapters, low-quality bases, and short reads.
3. **Dereplication**: Identical sequences are collapsed into unique sequences using tools like USEARCH (Edgar, 2010) or VSEARCH (Rognes *et al.*, 2016), reducing computational demands.

4. **Removal of Chimeras**: Chimera detection tools, such as UCHIME (Edgar *et al.*, 2011), identify and remove chimeric sequences that arise during PCR amplification.

5. **Contamination Filtering**: Non-target sequences, including human DNA and other environmental contaminants, can be filtered out using BLAST (Altschul *et al.*, 1990).

These steps remove erroneous reads and result in clustered sequences based on a predetermined similarity threshold. The clustered sequences are called Operational Taxonomic Units (OTUs) and are proxies for species-level identification in metabarcoding studies. The process of taxonomic annotation involves clustering sequences assigning the OTUs to known taxa using reference databases. Two primary approaches are used to group sequences; OTU Clustering, in which sequences are grouped based on similarity thresholds (commonly 97%), and Amplicon Sequence Variants (ASVs), that are exact sequence variants that offer higher resolution than OTUs. The taxonomic annotation requires comparison of OTUs/ASVs to reference databases, usually with BLAST-Based approaches against curated databases, such as Silva (Quast *et al.*, 2013) for bacteria, PR2 for protists (Guillou *et al.*, 2013) and BOLD (Barcode of Life Data Systems) for fish and macrofauna, or classifier tools like RDP Classifier (Wang *et al.*, 2007) and QIIME2's naive Bayes classifier, which provide taxonomic assignments based on training datasets.

Functional annotation enables researchers to estimate the functional diversity of communities, providing insights into ecosystem processes. Some commonly applied software for functional annotations of prokaryotic taxa include PICRUSt2 (Douglas *et al.*, 2020), which predicts the functional potential of microbial communities based on 16S rRNA data, Tax4Fun (Aßhauer *et al.*, 2015), a similar to PICRUSt tool that predicts functional profiles using SILVA-based reference trees, and FAPROTAX (Louca *et al.*, 2016), which maps taxa to ecological functions, particularly for biogeochemical cycling. A comprehensive understanding of marine biocommunities requires integrating taxonomic and functional data. Tools like QIIME2 offer modular pipelines for taxonomic and functional analyses, while the online KEGG Mapper collection of integrated tools visualizes functional pathways. Combining taxonomic and functional diversity metrics provides insights into ecosystem stability, resilience, and services.

In the next chapters we provide a detailed protocol to analyse raw read data from the fastq stage to final taxonomic and functional classifications. This pipeline, with slight ad hoc modifications will be used in the downstream processing of the metabarcoding data produced during NEMO-Tools.

1. eDNA extraction from filters

The efficacy of three sampling techniques for Microbial Communities and Fish Biodiversity assessment is featured in the context of the NEMO-Tools project. These sampling techniques included vacuum pump filtering, inline filtering, and passive samplers. Vacuum pump filtering (VP) offers unmatched efficiency for processing large water volumes, crucial for capturing eDNA in diverse and expansive marine environments. Its ability to handle substantial sample sizes enhances detection probabilities, making it ideal for comprehensive biodiversity assessments (Takahashi *et al.*, 2020). Inline filtering (Merck Millipore Sterivex filter units; IF), on the other hand, provides a streamlined and contamination-resistant option for fieldwork, preserving eDNA integrity from collection to laboratory analysis. Its portability and ease of use make it a practical choice for studies conducted in remote or resource-limited settings (Takahashi *et al.*, 2020). Lastly, passive samplers (rolls of gauze; PS) introduce a low-cost, low-labor alternative that captures eDNA over time, offering valuable temporal resolution and reducing the need for extensive field equipment (Maiello *et al.*, 2022). These three methods could provide a versatile toolkit that can be tailored to various marine research needs.

Environmental DNA was extracted using two commercially available kits; VP filters were treated with a Macherey—Nagel NucleoSpin® Soil Genomic DNA Isolation Kit (Qiagen), according to the manufacturer's instructions, whereas eDNA from IF and PS samples was extracted using a DNeasy® PowerWater® Genomic DNA Isolation Kit (Qiagen), according to the manufacturer's instructions. The concentration and quality of recovered DNA was confirmed using the Thermo Scientific™ NanoDrop™ spectrophotometer and an Invitrogen Qubit 4 Fluorometer.

2. Microbial Communities

The extracted DNA was subjected to PCR using specific primers targeting the V4 hyper variable region of the 18S rRNA gene (E572F = CYGCGGTAATTCCAGCTC; E1009R = AYGGTATCTRATCRTCTTYG) for the microeukaryotes (Comeau *et al.*, 2011), and the V3-V4 hyper variable regions of the 16S rRNA gene (S-DBact-0341-b-S-17: CCTACGGGNGGCWGCAG; S-D-Bact-0785-a-A-21: GACTACHVGGGTATCTAATCC) (Klindworth *et al.*, 2013) for the bacterioplankton. These primers have been found to successfully amplify approximately 450 - 460 bp. The amplicons will be sequenced on Illumina MiSeq using 300 + 300-bp paired-end chemistry, which allows for overlap and stitching together of paired amplicon reads into one full-length read.

After reviewing the existing bioinformatic tools and integrating with the applied pipelines in the research team's laboratories, we decided on the steps presented below to best clean the raw read datasets and produce a rigorous OTUs taxonomic and functional atlas of microbial species. We propose that the software mothur (https://mothur.org/) can best incorporate multiple tools into one command line and is one of the most time and computational efficient tools to handle large metabarcoding outputs.

3.1. Bioinformatic downstream cleaning of raw reads and clustering

The produced sequencing raw reads in fastq files will be subjected to downstream processing using the *mothur* v.1.48.0 software, following the proposed standard operating procedure (Schloss *et al.*, 2011). Before implementing the mothur SOP, quality checks will be performed with FastQC and low-quality reads or samples with overall low quality will be removed from further processing. The state-of-the-art high throughput sequencing chemistry and MiSeq machinery that is used in NEMO-Tools WP2 very rarely produces large output of erroneous data that could compromise entire samples. Thus, the probability to remove a sample because of low FastQC scores is very low considering the sampling, DNA extraction, and sequencing strategies we chose to follow. Within the mothur SOP, forward and reverse reads will be joined, and contigs below 200 bp, with >8 bp homopolymers and ambiguous base calls will be removed from further analysis, as the amplicon markers we target are expected to be around 450 bp for both prokaryotes and microeukaryotes. The remaining reads will be dereplicated to the unique sequences and aligned independently against the SILVA 132 database, containing eukaryotic and prokaryotic SSU rRNA gene sequences (Quast *et al.*, 2013). Then, the reads suspected of being chimeras will be removed using the UCHIME software (Edgar, 2010) which is embedded in the mothur command prompt. The remaining reads will be clustered into Operational Taxonomic Units (OTUs) at 97% similarity level (Behnke *et al.*, 2011), using the average neighbor method in mothur. To obtain a rigorous dataset, OTUs with a single read in the entire dataset will be removed from the analysis, as these are suspected of being erroneous sequences according to similar studies (e.g. Genitsaris *et al.*, 2016).

3.2. Taxonomic annotations

Taxonomic classification will be assigned using SINA searches against the SILVA database for bacteria (Pruesse *et al.*, 2012) and the mothur classify command against the Protist Ribosomal Reference database (PR2 database) with curated protistan taxonomy (Guillou *et al.*, 2013) for unicellular eukaryotes by applying the lowest accepted level of >80% similarity with a closest relative. The reads belonging to OTUs related to metazoa and land plants will be removed from the 18S rDNA dataset, since they do not belong to protists.

3.3. Prediction of functional diversity

To conduct the functional annotation of 16S rRNA gene amplicon Operational Taxonomic Units (OTUs) the PICRUSt tool will be used. A comprehensive pipeline is required that ensures the accurate functional profiling of the marine bacterial communities. The taxonomically annotated OTU table will be used as input for PICRUSt, which predicts functional profiles based on the abundance of OTUs and their corresponding phylogenetic placements. The first step in using PICRUSt involves normalizing the OTU table by correcting for 16S rRNA copy number, a step critical for ensuring the accuracy of downstream functional predictions. Once normalized, the PICRUSt pipeline proceeds to predict metagenome functions by mapping OTUs to their respective Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologs. This step generates a functional profile representing the potential metabolic capabilities of the microbial community. The predicted functional profile is analyzed to identify

key pathways and gene functions, with a focus on those relevant to aquatic ecosystems. The results can be further visualized and interpreted using tools like STAMP or statistical methods to compare functional profiles across samples or environmental gradients. Finally, validation of the functional predictions, whenever possible, should be conducted by comparing them to metagenomic data from the same samples, thus ensuring the robustness and reliability of the annotations provided by the PICRUSt tool. This pipeline offers a structured approach to integrate taxonomic and functional insights into the study of aquatic bacterial communities.

On the other hand, unicellular eukaryotes may belong to the same phylogenetic/taxonomic group but exhibit different functions. Based on this idea, the strategy that will be applied is to individually examine all produced OTUs of the dataset and, according to the taxonomic affiliations given by the PR2 database BLASTN search, proceed to literature searches concerning the biological descriptions for each taxon/OTU. Each OTU will then be annotated to specific biological trait sets for different functional groups (following Genitsaris *et al.* 2022 and Ramond *et al.* 2019) that were associated with resource acquisition, predator avoidance, and survival mechanisms (Irwin & Finkel, 2017). Usually, marine microeukaryotic OTUs are mainly affiliated to dinoflagellates, diatoms, ciliates, haptophytes, cryptophytes, and chlorophytes. The life strategies of these groups are well documented in the literature, and their functional annotations can be achieved with a high level of confidence. The rest of the OTUs, usually affiliated to taxonomic groups impossible to detect with microscopy, will be considered at the lowest taxonomic level available, and their functional annotations will be performed according to relevant literature. For example, OTUs belonging to the group of MArine ALveolates (MALV) will be considered parasites (Skovgaard, 2014), whereas the OTUs of MArine STramenopiles (MAST) will be grouped to nano-grazers (Massana *et al.* 2006), and OTUs associated with certain species of dinoflagellates, cryptophytes, and radiolarians will be considered mixotrophs based on relevant publications. These OTUs are reported to have both photosynthetic capacity and the ability to utilize heterotrophic strategies (see Genitsaris *et al.*, 2022). For traits associated with predator avoidance and survival, the Ocean Biodiversity Information System (OBIS; https://obis.org/) and the Global Biodiversity Information Facility (GBIF; www.gbif. org/) will be additionally consulted. The OTUs belonging to taxa with parasitic lifestyles will be annotated as potentially harmful and the IOC-UNESCO Taxonomic Reference List of Harmful Microalgae will be used to determine potential harmfulness for the remaining OTUs. Finally, the Protist Interaction Database (PIDA) on the planktonic protist interactome was also consulted (Bjorbækmo *et al.* 2020).

For a summarized step-by-step processing of the produced reads into taxonomic and functional identification of the different microbial biocommunities, from the fastq step to the production of OTUs tables and predicted functional occurrences, see Table 1.

Table 1. Command lines in mothur and PICRUSt to taxonomic and functional endpoints.

Commands	Description
fastQC quality check	Checks the quality of produced fastq files after high-throughput sequencing
make.file	Creates a description txt file that summarizes the fastq files

make.contigs	Merges the forward and reverse reads into one contig per sequence
screen.seqs	Removes reads with ambiguities and non-expectant sizes
unique.seqs	Decreases the size of the produced fasta file by keeping only one copy per reads with 100% similarity between them
align.seqs	Aligns reads based on the curated Silva database
filter.seqs	Removes tabs and dots from the aligned reads
pre.cluster	Groups reads at a preliminary clustering level permitting one polymorphism per 100 bases
chimera.vsearch	Uses the vsearch software to identify and remove potential chimeric reads
split.abund	Further decreases the size of the produced fasta by separating reads with only one occurrence in the entire dataset. These are suspected to be erroneous reads
dist.seqs	Gathers reads with a cutoff = 0.03. Essentially prepares the dataset for the final clustering into OTUs with >97% sequence similarity levels
cluster	Clusters reads with >97% similarity into OTUs
make.shared	Creates a datamatrix with the number of reads per OTU per sample
get.oturep	Gets the final fasta file
picrust2_pipeline.py	Assigns predicted functional and pathway abundances

3. Fish

The extracted eDNA will be subjected to PCR using the fish-specific MiFish primer pair targeting a hypervariable region of the mitochondrial 12S rRNA gene (F: 5'-GTCGGTAAAACTCGTGCCAGC-3'; R: 5'- CATAGTGGGGTATCTAATCCCAGTTTG -3'; Miya *et al.*, 2015). These primers have been found to successfully amplify approximately 172 bp. The amplicons will be sequenced on Illumina 6000 using 2 x 150-bp paired-end chemistry, which allows for overlap and stitching together of paired amplicon reads into one full-length read.

After reviewing a number of existing bioinformatic tools and pipelines, we decided to use the MJOLNIR pipeline (https://github.com/uit-metabarcoding/MJOLNIR/) that can handle raw read metabarcoding datasets and transform them into organized data sets of taxonomically assigned MOTUs (Fig. 1), using the developer's optimization suggestions.

4.1. Bioinformatic downstream analysis of raw metabarcoding reads

The produced sequencing raw reads in multiplexed fastq files are processed through the MJOLNIR pipeline (https://github.com/uit-metabarcoding/MJOLNIR), primarily utilizing tools from the OBITools package (Boyer *et al.*, 2016).

1) In the first step (RAN), multiplexed FASTQ files are split into aliquot parts for parallel processing.

2) Afterwards, FREYJA includes three steps: paired-end alignment, demultiplexing, and read-length filtering. Paired-end reads are aligned using *illuminapairedend*, retaining only those with an alignment quality score above 40. Demultiplexing and primer-sequence removal are performed by *ngsfilter*, discarding sequences with mismatched primer tags. Length filtering and dereplication of sequences are performed with *obigrep* and *obiuniq*, retaining sequences of 140-190 bp. Additionally, reads containing bases apart from [ACGT] are also erased.

3) During HELA, singleton sequences and chimeric amplicons are eliminated using the *uchime_denovo* algorithm from VSEARCH (Rognes *et al.*, 2016). After removing chimaeras from samples, HELA merges samples again, and de-replicates all sequences. Two output files are produced: a fasta file with total abundances for clustering, and a table with abundance information of all sequence variants per sample.

4) During ODIN, molecular operational taxonomic units (MOTUs) are delimitated based on linkage networks created by step-by-step aggregation, using the SWARM procedure (Mahé *et al.*, 2015).

4.2. Taxonomic annotations and reference database

Taxonomic assignment is performed in step 5) THOR, which is a wrapper function of *ecotag* (Boyer *et al.* 2016) and *owi_add_taxonomy* (Wangensteen & Turon 2017). For the annotation, we used the DNA Universal-databank for Fisheries and Aquaculture reference database (DUFA; last updated on 2022-01-06) for the 12S MiFish fragment (https://github.com/uit-metabarcoding/DUFA). Additionally, DUFA's completeness for non-indigenous species in the Mediterranean Sea was cross-referenced with an updated list (Zenetos *et al.*, 2022). We manually added 79 published 12S sequences absent in the DUFA 12S database (excluding unverified entries). These sequences were retrieved from the National Center for Biotechnology Information (NCBI) Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy). After the taxonomic assignment,

6) FRIGGA connects the taxonomic information acquired during the previous step with the information of abundances per sample calculated in step 4 (ODIN), and during

7) LOKI, putative pseudogenes are removed using LULU (Frøslev *et al.*, 2017). In the final step,

8) RAGNAROC produces a final community dataset with MOTU's read count per sample and the taxonomic information of each assigned MOTU.

4.3. Post-bioinformatic community data manipulation

For community analysis, environmental triplicate samples will be merged and treated as a single sample. A minimum of 97% identity will be used for species level identification (Miya *et al.*, 2015). Additionally, MOTUs with relative read abundance per sample below 0.005 % or less than five reads will be set to zero to minimize tag-switching bias (Antich *et al.*, 2023) and reduce the likelihood of false positives from potential contamination. All datasets will be treated as qualitative presence-absence data on each MOTU per sample.

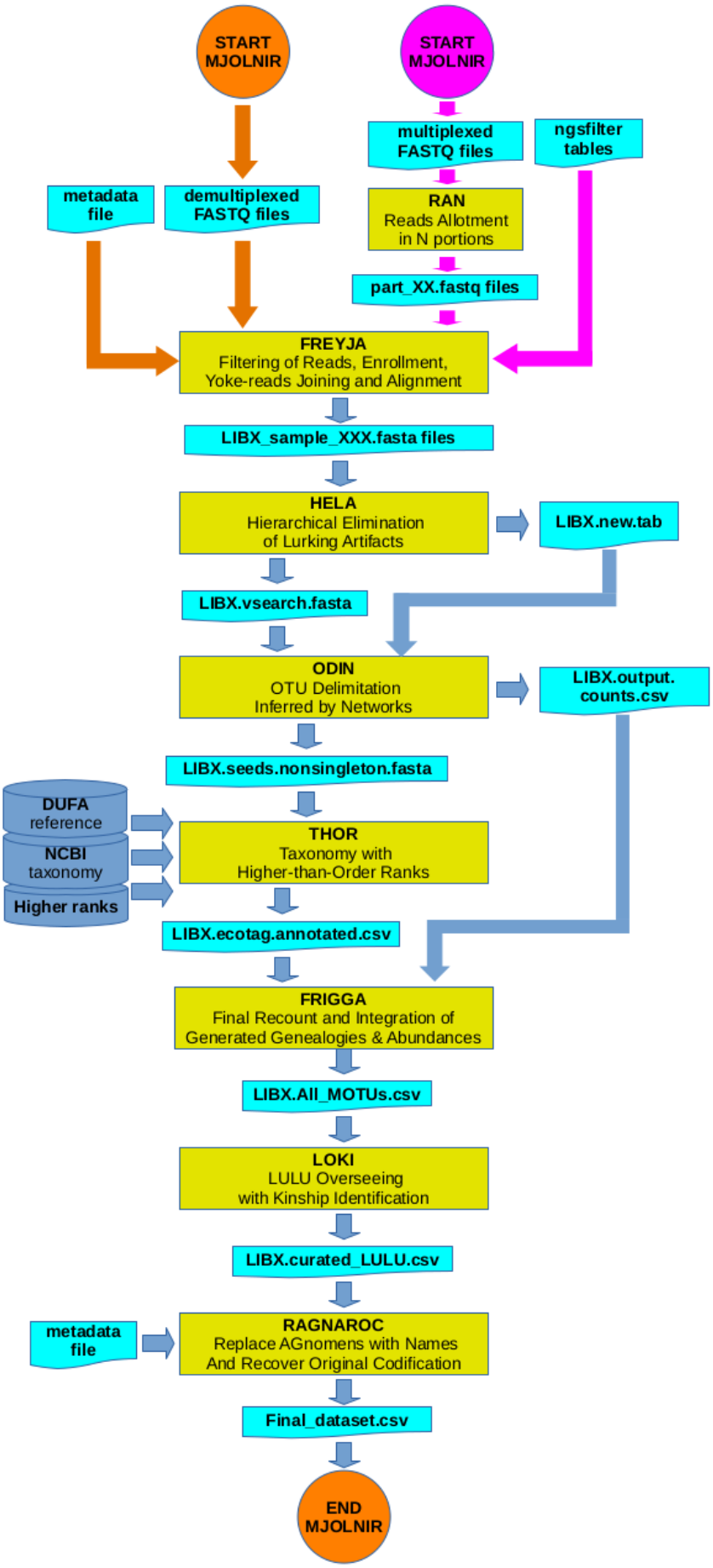
START MJOLNIR
START MJOLNIR
multiplexed FASTQ files
ngsfilter tables
metadata file
demultiplexed FASTQ files
RAN
Reads Allotment in N portions
part_XX.fastq files
FREYJA
Filtering of Reads, Enrollment, Yoke-reads Joining and Alignment
LIBX_sample_XXX.fasta files
HELA
Hierarchical Elimination of Lurking Artifacts
LIBX.new.tab
LIBX.vsearch.fasta
ODIN
OTU Delimitation Inferred by Networks
LIBX.output. counts.csv
LIBX.seeds.nonsingleton.fasta
DUFA reference
NCBI taxonomy
Higher ranks
THOR
Taxonomy with Higher-than-Order Ranks
LIBX.ecotag.annotated.csv
FRIGGA
Final Recount and Integration of Generated Genealogies & Abundances
LIBX.All_MOTUs.csv
LOKI
LULU Overseeing with Kinship Identification
LIBX.curated_LULU.csv
metadata file
RAGNAROC
Replace AGnomens with Names And Recover Original Codification
Final_dataset.csv
END MJOLNIR

Figure 1. The MJOLNIR pipeline workflow (https://github.com/uit-metabarcoding/MJOLNIR/) starting with raw metabarcoding data and finishing with a final community dataset.

4. Challenges of eDNA

While eDNA metabarcoding has advanced biodiversity studies, challenges remain:

- **Database Limitations**: Reference databases often lack comprehensive coverage of marine taxa, particularly for understudied regions, cryptic species, and non-model organisms.
- **Functional Prediction Accuracy**: Tools like PICRUSt2, rely on phylogenetic assumptions that may not hold true across all taxa.
- **Environmental Variables**: Linking functional diversity to environmental gradients requires integrating eDNA data with abiotic and biotic factors.

Future advancements in high-throughput sequencing, bioinformatics, and machine learning will likely address these challenges, enabling more precise and comprehensive analyses of marine microbial diversity.

5. Perspectives

The comprehensive protocol developed for NEMO-Tools represents a crucial advancement in the taxonomic and functional identification of marine microbial communities. By employing high-throughput sequencing (HTS) technologies and a meticulous bioinformatics pipeline, this deliverable provides a step-by-step methodology to process raw sequencing reads into meaningful ecological insights. Specifically, the protocol leverages tools like FastQC for quality assessment, and mothur/mjolnir for trimming, quality filtering and clustering sequences into Operational Taxonomic Units (OTUs) at a 97% similarity threshold. These OTUs serve as proxies for species-level identification, facilitating downstream taxonomic annotation through curated reference databases such as SILVA for bacterial species, the PR2 database for protists, and the DUFA for fishes. This systematic approach ensures that erroneous reads, artifacts, and contaminants are filtered out, leading to reliable and reproducible taxonomic datasets. Functional annotation is a cornerstone of this protocol, employing PICRUSt2 to predict metagenomic functions from 16S rRNA gene data. By normalizing OTU tables and mapping them to KEGG orthologs, the protocol generates detailed profiles of microbial functional potential, highlighting their roles in nutrient cycling, biogeochemical processes, and ecosystem stability. For

microeukaryotic taxa, the pipeline goes further by incorporating literature-based trait assignments to categorize OTUs into functional groups related to resource acquisition, predator avoidance, and survival mechanisms. This dual focus on taxonomic and functional data enables a holistic understanding of microbial communities and their ecological roles within marine ecosystems. A key strength of the protocol lies in its adaptability to varying sample types and environmental conditions. For example, it accommodates different levels of taxonomic resolution, from broad OTU clustering to exact sequence variants (ASVs), depending on the research question and data quality. Additionally, the integration of multiple databases, including the IOC-UNESCO Harmful Algae List and the Protist Interaction Database, ensures comprehensive annotations even for rare or poorly studied taxa. Challenges, such as incomplete database coverage and the inherent limitations of predictive tools like PICRUSt2, are acknowledged and addressed through complementary validation strategies, including comparison with metagenomic datasets.

The bioinformatics pipeline of macroeukaryotes, particularly of fishes, can also be challenging. Nevertheless, there are key point that affect community assessment. The choice of molecular markers for meta-barcoding undeniably influences the outcome of eDNA studies (Zangaro *et al.*, 2021; Zhang *et al.*, 2020). For microeukaryotes, the PR2 database offers a large primer database which can provide coverage information of frequently used set of primers (Vaulot *et al.*, 2022). Mediterranean fish, as a taxonomic group, have a considerably extensive barcode reference database, mostly referring to the cytochrome c oxidase I (COI) gene (Zangaro *et al.*, 2021) but this marker is not as suitable for fish detection and monitoring. Moreover, there are still species that are absent from public databases, particularly for other barcoding regions, e.g., 12S rRNA. For the NEMO-Tools project, we selected the MiFish primers pair due to its higher fish specificity and its overall improved performance in fish detection to other primers. For example, it produces greater number and proportion of usable fish reads to COI primers despite the more complete reference database of the latter primer pair (Collins *et al.*, 2019; Zhang *et al.*, 2020). Despite the increased use of eDNA techniques for rare species detection, the incompleteness of reference databases is the most restricting factor in monitoring and management for the rare microbial biosphere and the rare and endangered fish in the Mediterranean Sea. Moreover, primer specificity and efficiency for targeting specific groups should be considered. Numerus assays with group- or species-specific primers are designed, promising higher levels of precision and efficiency (Ardura, 2019; Hartle-Mougiou *et al.*, 2024). Additionally, the combined use of multiple primers could enhance species detection and/or richness as multi-marker approaches often yield more detailed and higher resolution community data by targeting more taxonomic groups (Ferreira *et al.*, 2024; Fontes *et al.*, 2024).

Protocol optimization could focus on DNA yield by testing filtered seawater volume and/or using inhibitor removal kits. Finding the optimal volume is extremely beneficial to monitoring strategies, particularly as small volumes can be advantageous due to their minimal physical and logistical requirements, faster sampling, processing times,

and equipment simplicity. However, higher water volumes can improve eDNA metabarcoding results and consistency (Bessey *et al.*, 2020; Govindarajan *et al.*, 2022). Additionally, different extraction methods, types of filters, environmental parameters, and protocol costs should be examined (Duarte *et al.*, 2021 Rishan *et al.*, 2023). Finally, a sufficient sample size should provide a stronger basis for statistical analysis, improving our ability to assess the impact of environmental variables on microbial diversity retrieval and threatened fish detection. While eDNA methods can be effective for biodiversity monitoring, a homogenized protocol standardization is required to maximise their potential.

In total, the applied protocols not only set a benchmark for the NEMO-Tools project but also contributes to the broader field of marine ecology by offering a standardized and scalable framework. Its application will enable researchers to track microbial and fish diversity and functionality across spatial and temporal scales, providing critical data to inform conservation and management strategies. As marine ecosystems face mounting pressures from climate change and anthropogenic activities, the ability to assess community dynamics with precision is more important than ever.

6. References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J., 1990. Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Antich, A., Palacín, C., Zarcero, J., Wangensteen, O. S., Turon, X., 2023. Metabarcoding reveals high-resolution biogeographical and metaphylogeographical patterns through marine barriers. *Journal of Biogeography*, 50(3), 515-527.

Ardura, A., 2019. Species-specific markers for early detection of marine invertebrate invaders through eDNA methods: Gaps and priorities in GenBank as database example. *Journal for Nature Conservation*, 47, 51–57, doi: 10.1016/j.jnc.2018.11.005

Aßhauer, K. P., Wemheuer, B., Daniel, R., Meinicke, P., 2015. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics*, *31*(17), 2882-2884.

Behnke, A., Engel, M., Christen, R., Nebel, M., Klein, R. R. *et al.*, 2011. Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environmental microbiology*, *13*(2), 340-349.

Bessey, C., Jarman, S. N., Berry, O., Olsen, Y. S., Bunce, M. *et al.*, 2020. Maximizing fish detection with eDNA metabarcoding. *Environmental DNA*, 2(4), 493-504.

Bjorbækmo, M. F. M., Evenstad, A., Røsæg, L. L., Krabberød, A. K., Logares, R., 2020. The planktonic protist interactome: where do we stand after a century of research?. *The ISME journal*, *14*(2), 544-559.

Bolger, A. M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C. *et al.*, 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature biotechnology*, *37*(8), 852-857.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. *et al.*, 2016. obitools: A unix-inspired software package for DNA metabarcoding. *Molecular ecology resources*, 16(1), 176-182.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A. *et al.*, 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature methods*, *13*(7), 581-583.

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L. *et al.*, 2019. Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, 10(11), 1985-2001.

Comeau, A. M., Li, W. K., Tremblay, J. É., Carmack, E. C., Lovejoy, C., 2011. Arctic Ocean microbial community structure before and after the 2007 record sea ice minimum. *PloS one*, *6*(11), e27492.

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R. *et al.*, 2020. PICRUSt2 for prediction of metagenome functions. *Nature biotechnology*, *38*(6), 685-688.

Duarte, S., Vieira, P. E., Lavrador, A. S., Costa, F. O., 2021. Status and prospects of marine NIS detection and monitoring through (e) DNA metabarcoding. *Science of the Total Environment*, 751, 141729.

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

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194-2200.

Ferreira, A. O., Azevedo, O. M., Barroso, C., Duarte, S., Egas, C. *et al.*, 2024. Multi-marker DNA metabarcoding for precise species identification in ichthyoplankton samples. *Scientific Reports*, 14(1), 19772.

Fontes, J. T., Katoh, K., Pires, R., Soares, P., Costa, F. O., 2024. Benchmarking the discrimination power of commonly used markers and amplicons in marine fish (e) DNA (meta) barcoding. *Metabarcoding and Metagenomics*, 8, e128646.

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., *et al.*, 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, 8, 1188.

Genitsaris, S., Monchy, S., Breton, E., Lecuyer, E., Christaki, U., 2016. Small-scale variability of protistan planktonic communities relative to environmental pressures and biotic interactions at two adjacent coastal stations. *Marine Ecology Progress Series*, *548*, 61-75.

Genitsaris, S., Stefanidou, M., Sommer, U., Moustaka-Gouni, M., 2022. Diversity of taxon-specific traits of seasonally distinct unicellular eukaryotic assemblages in a eutrophic coastal area with marked plankton blooms. *Aquatic Microbial Ecology*, *88*, 167-185.

Goodwin, S., McPherson, J. D., McCombie, W. R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature reviews genetics*, *17*(6), 333-351.

Govindarajan, A. F., McCartin, L., Adams, A., Allan, E., Belani, A. *et al.*, 2022. Improved biodiversity detection using a large-volume environmental DNA sampler with in situ filtration and implications for marine eDNA sampling strategies. *Deep Sea Research Part I: Oceanographic Research Papers*, 189, 103871.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C. *et al.*, 2013. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic acids research*, *41*(D1), D597-D604.

Hartle-Mougiou, K., Gubili, C., Xanthopoulou, P., Kasapidis, P., Valiadi, M. *et al.*, 2024. Development of a quantitative colorimetric LAMP assay for fast and targeted molecular detection of the invasive lionfish Pterois miles from environmental DNA. *Frontiers in Marine Science*, 11, 1358793.

Irwin, A. J., Finkel, Z. V., 2017. Phytoplankton functional types: a trait perspective. *BioRxiv*, 148312.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C. *et al.*, 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic acids research*, *41*(1), e1-e1.

Louca, S., Parfrey, L. W., Doebeli, M., 2016. Decoupling function and taxonomy in the global ocean microbiome. *Science*, *353*(6305), 1272-1277.

Mahé, F., Rognes, T., Quince, C., De Vargas, C., Dunthorn, M., 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420.

Maiello, G., Talarico, L., Carpentieri, P., De Angelis, F., Franceschini, S. *et al.*, 2022. Little samplers, big fleet: eDNA metabarcoding from commercial trawlers enhances ocean monitoring. *Fisheries Research*, *249*, 106259.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), 10-12.

Massana, R., Terrado, R., Forn, I., Lovejoy, C., Pedrós-Alió, C., 2006. Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. *Environmental microbiology*, *8*(9), 1515-1522.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y. *et al.*, 2015. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. *Royal Society open science*, 2 (7), 150088.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T. *et al.*, 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic acids research*, *41*(D1), D590-D596.

Reuter, J. A., Spacek, D. V., Snyder, M. P., 2015. High-throughput sequencing technologies. *Molecular cell*, *58*(4), 586-597.

Rishan, S. T., Kline, R. J., Rahman, M. S., 2023. Applications of environmental DNA (eDNA) to detect subterranean and aquatic invasive species: a critical review on the challenges and limitations of eDNA metabarcoding. *Environmental Advances*, 12, 100370.

Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F., 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, 4, e2584.

Schloss, P. D., Gevers, D., Westcott, S. L., 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PloS one*, *6*(12), e27310.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M. *et al.*, 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, *75*(23), 7537-7541.

Skovgaard, A., 2014. Dirty tricks in the plankton: diversity and role of marine parasitic protists. *Acta Protozoologica*, *53*(1).

Stat, M., Huggett, M. J., Bernasconi, R., DiBattista, J. D., Berry, T. E. *et al.*, 2017. Ecosystem biomonitoring with eDNA: metabarcoding across the tree of life in a tropical marine environment. *Scientific reports*, *7*(1), 12240.

Takahashi, S., Sakata, M. K., Minamoto, T., Masuda, R., 2020. Comparing the efficiency of open and enclosed filtration systems in environmental DNA quantification for fish and jellyfish. *PLoS One*, *15*(4), e0231718.

Thomsen, P. F., Willerslev, E., 2015. Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. *Biological conservation*, *183*, 4-18.

Vaulot, D., Geisen, S., Mahé, F., Bass, D., 2022. pr2-primers: An 18S rRNA primer database for protists. *Molecular Ecology Resources*, *22*(1), 168-179.

Wang, Q., Garrity, G. M., Tiedje, J. M., Cole, J. R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology*, *73*(16), 5261-5267.

Wangensteen, O. S., Turon, X., 2017. Metabarcoding techniques for assessing biodiversity of marine animal forests. *Marine animal forests. The ecology of benthic biodiversity hotspots*, *1*, 445-503.

Zangaro, F., Saccomanno, B., Tzafesta, E., Bozzeda, F., Specchia, V. *et al.*, 2021. Current limitations and future prospects of detection and biomonitoring of NIS in the Mediterranean Sea through environmental DNA. *NeoBiota*, 70, 151-165.

Zenetos, A., Albano, P. G., Garcia, E. L., Stern, N., Tsiamis, K. *et al.*, 2022. Established non-indigenous species increased by 40% in 11 years in the Mediterranean Sea. *Mediterranean Marine Science*, 23(1), 196-212.

Zhang, S., Zhao, J., Yao, M., 2020. A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution*, 11 (12), 1609–1625.