Development of a User-Friendly Graphical Platform for Molecular Characterization of Macroscopic Parasites in Fish

João Faria[1], João Carneiro[2,3], and Teresa Rito[2,3]

[1] Informatics Department, University of Minho, Braga, Portugal
[2] Centre of Molecular and Environmental Biology (CBMA), Department of Biology, University of Minho, Braga, Portugal
[3] Institute of Science And Innovation for Bio-Sustainability (IB-S), University of Minho, Braga, Portugal

1 Introduction

Parasites are ubiquitous parts of biological systems, accounting for a sizable portion of global biodiversity. They are among nature's most successful life forms, having emerged repeatedly independently in numerous phyla [1]. In aquatic environments, parasites' prevalence and evolutionary success are especially noticeable because of their complex interactions with fish. They can change factors that are frequently employed in fish ecology and fisheries, and they may oppose or complement host parameters that are typically utilized as markers of physiological or reproductive health. Parasite-induced alterations in host behavior can also alter patterns of host distribution, habitat choice, diet composition, sexual behavior, and others, which can have an impact on fish ecology as well as that of their predators and prey [1].

2 State of the Art

Given the significant role of parasites in aquatic systems, accurate and efficient identification methods are crucial for understanding their impact. Advancements in sequencing technologies, which enable the generation of tens to hundreds of millions of reads in a single sequencing run, have facilitated the development of novel research methodologies. These approaches encompass species mapping, biomonitoring, gut content analyses, and population genomics, thereby significantly enhancing studies of marine ecosystems [2]. The advantages of metabarcoding have been recognized by parasitologists and other academics. Due to this, metabarcoding has become much more popular in the past ten years for applications such as helminth parasite detection. Aivelo and Medlar [3] evaluated four helminth metabarcoding studies to determine the method's applicability in parasitological research along with its advantages and disadvantages. These authors found that metabarcoding allowed for non-invasive sampling techniques while offering a quick and thorough examination of the composition of parasite communities [4]. High-throughput multi-species identification was initially referred to

as DNA metabarcoding by Teberlet [5]. Next-generation sequencing (NGS) technology is pivotal to this process, as it provides the massive parallel sequencing capacity required to analyze environmental samples or bulk collections of complete organisms. Even though traditional DNA barcoding simply streamlines the taxonomic process of creating a list of species found in an environment and does little to minimize sampling effort, the introduction of NGS presents an opportunity to overcome this limitation. If we accept that all living organisms leave traces of their DNA in the environment, it becomes possible to gather this environmental DNA, use PCR to amplify barcode markers from it, and then sequence the resulting amplicons using NGS [6]. Essentially, NGS makes DNA metabarcoding feasible by rapidly generating vast amounts of sequence data, which allows for the simultaneous detection of multiple species, even when present in trace amounts. This method provides a cost-effective and non-invasive way to monitor organisms while minimizing the need for in-depth taxonomic expertise and reducing reliance on often ambiguous morphological traits [7]. Thus, by enhancing detection sensitivity and expediting multi-species identification processes, the integration of NGS with DNA metabarcoding significantly improves biodiversity assessments. To effectively use such advantages, though, they must be coupled with detailed and carefully chosen reference libraries, which serve as the foundation of accurate species identification. The development and widespread application of DNA metabarcoding for biomonitoring and biodiversity assessments has significantly increased the demand for high-quality reference libraries. Due to the huge volume of reads from high-throughput sequencing (HTS) tools, automated systems that compare query sequences to reference sequences in DNA sequence libraries like National Center for Biotechnology Information's (NCBI) GenBank or the Barcode of Life Data Systems (BOLD) are frequently part of essential bioinformatics. Despite these promising benefits, several limitations persist in optimizing metabarcoding workflows. In some cases, new approaches are now incorporating the supervision of quality control within reference data collection, ensuring greater accuracy and reliability in the results [8, 9]. Erroneous data in reference libraries can cause repeated and undetected identification errors over time [10]. To ensure that this project incorporates the most recent advancements in methodologies and technologies, Elicit AI, a literature review search engine [11] powered by large language models (LLMs), was utilized to facilitate the identification of current innovations in metabarcoding techniques and the curation of reference databases. These challenges highlight the critical need for improved bioinformatic tools tailored specifically to fish parasite detection. It is crucial for varied populations like fish that reference databases are complete. The biodiversity of marine ecosystems present challenges for manual database curation, unlike the relatively confined and well-documented nature of freshwater habitats. Recently, the use of environmental DNA (eDNA) metabarcoding has emerged as a promising methodological approach for the study of marine fish populations, demonstrating significant potential as an alternative tool for ecosystem monitoring. In this regard, using eDNA-based techniques in fisheries monitoring would depend heavily on the availability of comprehensive and main-

tained datasets. Universal metazoan cytochrome oxidase subunit I (COI) primers often amplify non-target species due to the abundance of eDNA in water samples. Therefore, fish-specific primers like MiFish have been developed [12].

3 Problem Statement

A specific tool called PMiFish has been developed to further improve the analysis of fish-specific eDNA data collected with MiFish primers. The program, which was developed with the language Perl, combines MEGA X for creating phylogenetic trees with USEARCH v11 for quick sequence analysis, enabling researchers to summarize and analyze results. However, it may not be user-friendly for everyone as it often requires a Linux-based environment [13]. Despite these advances, metabarcoding analysis still faces several obstacles. The process involves managing complex sequencing data flows, curating reference databases such as those for fish parasites, and integrating various bioinformatic modules into a unified system. These tasks present challenges for existing procedures.

4 Objectives

The study aims to develop a user-friendly, cross-platform graphical interface for simplifying raw data processing. This includes constructing phylogenetic trees, identifying species using curated databases of fishes and fish parasites, cataloging, and merging samples to address these issues. Our approach will allow non-specialists to do complex molecular analysis by adding a rich reference library of fish parasite sequences and improving the PMiFish pipeline with an intuitive user interface. The primary aim of this project is to minimize fish food waste by improving parasite detection and control. This will be achieved by improving bioinformatic techniques to increase the accuracy of sequence analysis. In summary, the project aims to revolutionize fish parasite detection and ecosystem monitoring by integrating cutting-edge bioinformatics tools, curated reference databases, and advanced sequencing technologies. By minimizing food waste, enhancing multi-species identification, and enabling accurate phylogenetic analyses, this initiative seeks to provide a comprehensive and user-friendly platform for researchers and professionals in the field of fisheries science and molecular ecology.

5 Methodology

Our methodology will start with the quality control of raw sequencing data using FQC, a web-based dashboard that incorporates metrics generated by FastQC into an extendable and interactive framework. FQC allows for the dynamic presentation of important parameters such as read abundance patterns, adaptor content, GC distribution, and per-base sequence quality through aggregating

quality control findings from several sequencing runs [14]. Then we will perform trimming with Trimmomatic to remove low-quality regions and adapter sequences [15]. To be able to rebuild full-length sequences for accurate downstream analysis, paired-end reads will then be merged with Paired-End reAd mergeR (PEAR) [16]. Similarly, single-end reads will be processed using FQC for quality control and Trimmomatic for trimming. The high-quality reads will then be used directly for downstream analyses, removing the need for a merging step and guaranteeing consistent quality filtering across both data types. Regardless of the initial format of the sequencing datasets, we will be able to conduct thorough analysis by preserving parallel workflows for both paired-end and single-end reads. Following this step, we will concentrate on understanding operational taxonomic units OTUs and amplicon sequence variants (ASVs) in their entirety and evaluating their influence on bioinformatics processes [17,18]. OTUs are characterized as a group of sequences with a sequence identity over a predetermined threshold, usually 97%. There are now many different OTU clustering algorithms that can cluster reads in a variety of ways and offer a limitless variety of parameter options for the strictness (or inclusivity) of the OTU clustering [17]. One example of such an algorithm is USEARCH [19,20], which is also used in the PMiFish pipeline. VSEARCH is a free, open-source, multithreaded 64-bit tool for processing nucleotide sequence data in metagenomics, genomics, and population genomics. It is an alternative to USEARCH, which offers only a closed-source, limited, 32-bit version for academic use [21]. Based on these findings, we will implement VSEARCH for OTU generation within our platform. However, with the rise of denoising methods, there has been a shift towards ASVs instead of clusters. These methods use sequencing quality to differentiate true variations from errors, identifying distinct ASVs even when there's a single nucleotide difference. At the moment, DADA2 and Deblur are the two most often employed programs for ASV calculation [17], and when compared to Deblur, DADA2 typically generates more Amplicon Sequence Variants (ASVs) [22]. DADA2, an open-source R package, transforms and enhances the DADA algorithm [23] by employing a parametric model to deduce authentic biological sequences from the data. True reads are likely to be more plentiful, and the model depends on input read abundances and distances; less abundant reads, which are only a few base differences away from a more abundant sequence, are likely error-derived [19]. These results, which are supported by the fact that DADA2 is the pipeline most frequently used among researchers [4], led us to select DADA2 as the ASV analysis software, which will be implemented on our platform. In addition to these clustering and denoising methods, rapid taxonomic classification remains essential, prompting the integration of direct classification tools. Kraken2 is a sophisticated taxonomy categorization program that uses a precise k-mer alignment method to classify sequences. It divides each query sequence into k-mers, matches them with a specific taxon via a hash table, and assigns a label based on the weighted average of these matches. Kraken2 is ideal for amplicon analysis and shotgun metagenomics due to its efficient database structure and classification speed [24,25]. Additionally, Kraken2 supports 16S rRNA

classification using databases such as Greengenes, SILVA, and RDP [26]. Thus, we will include an option in our software allowing researchers to use Kraken2 for fast analysis. Our platform will integrate both OTU and ASV approaches, along with Kraken2, for direct classification, providing flexibility for different experimental needs. Ensuring accurate species identification also requires careful selection and integration of reference databases. We will analyze each database's characteristics, including the availability of an Application Programming Interface (API), data volume, and the specific focus and specialization of its datasets, to determine the optimal integration strategy for our platform. With NCBI GenBank being the most widely used database, along with SILVA and PR2, which are also extensively utilized, researchers mostly rely on bioinformatics databases for sequence identification, as reported in [4]. When selecting a bioinformatic database, it is crucial to consider whether the genetic marker used is compatible with the sequences in the database. The gene encoding the small ribosomal subunit (18S rRNA) is the most commonly used molecular marker for eukaryotic microorganisms since it is ubiquitous and reference databases such as SILVA and PR2 are carefully curated [27]. Additionally, the COI gene and other ribosomal RNA genes (16S and 12S) are also essential markers for eukaryotic research [28]. Knowing these factors and considering that each database is particularly suited to a distinct set of markers, we will integrate them all into our platform. This integration will provide researchers with access to a broader range of analytical tools. Large-scale metabarcoding investigations have been undertaken using the 16S rRNA gene for bacteria and archaea [27, 29]. The SILVA database will be employed for these purposes since it is only compatible with rRNA markers like 16S and 18S, just like the PR2 database. Additionally, because they are smaller databases, nearly all of the sequences are annotated, and they are regularly updated and quality-checked [4]. Both rRNA genes and cytochrome c oxidase subunit I (COI) sequences are widely used for animal research. Because of its widespread application and crucial function in the BOLD database, COI in particular has become known as the preferred marker. Notably, there are two main repositories where COI data are kept. The first is the NCBI-nt, a generalist nucleotide database, which includes a variety of taxa and genes, alongside its European (ENA) and Japanese (DDBJ) counterparts. The second repository is the BOLD, which primarily holds the COI gene's barcoding segment even though it also contains sequences for other markers [30]. The COI animal marker reference datasets from GenBank and BOLD are complementary to one another; however, since both databases are updated continuously, studies carried out within a short period may report differences of thousands of reference sequences [31]. With these findings in consideration, it is pertinent to note that GenBank continues to be the largest nucleotide sequence database [4], but its open submission policy promotes broad access to data as well as allows the accumulation of unconfirmed and possibly misannotated records. While GenBank has been acknowledged for its use in several DNA-based monitoring applications, its reliability has been questioned, with problematic entries reported in several taxa [12]. The BOLD database will be our primary source for COI gene information and analysis,

with NCBI GenBank used for cross-referencing to minimize annotation errors. Users can choose the appropriate database for their research, with features and restrictions clearly displayed. Smaller databases like SILVA and PR2 can be downloaded locally, while larger ones like NCBI will be accessed via APIs. To handle the large BOLD database efficiently, we will integrate the Python package, BOLDigger. The BOLD website offers access to both early-release and published COI records, which make up about half of all BOLD records. It has a limit of processing 100 sequences at a time. BOLDigger addresses this by automatically requesting up to 100 sequences, ensuring the limits are not exceeded and enhancing the determination of top hits [32]. Next, sequence alignment for phylogenetic analysis will be performed using MAFFT, recognized for its accuracy and runtime efficiency. FastTree 2 will be used under the GTR model to quickly generate a tree topology, followed by branch length optimization with RAxML-NG [33]. Our platform's user interface will be developed with Plotly's Dash, a library that allows data scientists to create declarative interactive web applications in Python [34], ensuring a cross-platform, user-friendly experience that simplifies complex bioinformatics workflows for non-specialist users. As detailed in our methodology, Figure 1 demonstrates the workflow of our metabarcoding platform, outlining each analytical stage from the initial acquisition of raw sequencing data to the final interactive visualization of results.

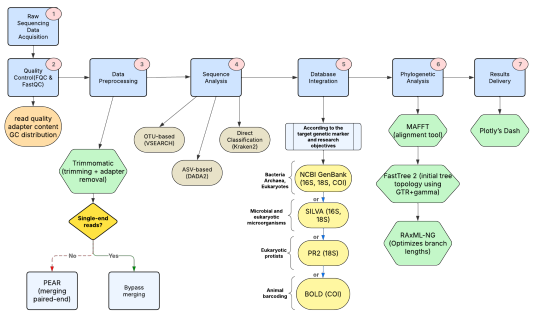

Fig. 1. Workflow diagram of the metabarcoding platform.

6 Expected Outcomes

The project is expected to produce a cross-platform, intuitive metabarcoding software solution that enhances PMiFish. The platform is anticipated to improve classification accuracy and simplify complex bioinformatics procedures by combining an easy-to-use interface and a carefully curated parasite sequencing

reference library. Enhancing fish parasite management will benefit aquaculture, researchers, and food safety regulators.

7 Results: PARAFISH Workflow

To demonstrate the capabilities and utility of the PARAFISH platform—named for the fact that it was developed for the analysis of fish parasite samples—we provide an in-depth explanation of a standard metabarcoding analytical methodology. An example dataset of fish parasite samples (COI marker) was utilized to demonstrate each step of the pipeline, from raw data upload to phylogeny inference, with a focus on significant outputs and their biological interpretation. Due to time constraints DADA2, Kraken2, and the integration of additional reference databases beyond BOLD, are not yet available in the present version. Though the present implementation focuses on offering a reliable and user-friendly platform, these updates are planned for future releases. Paired-end sequencing reads that target the mitochondrial cytochrome c oxidase I (COI) gene specifically constitute the dataset selected for this demonstration. Amplification of a fragment of the mitochondrial cytochrome c oxidase I (COI) gene was performed using the primers mlCOIintF (forward: GGWACWGGWTGAACWGTWTAYCCYCC) and LoboR1 (reverse: TAAACYTCWGGRTGWCCRAARAAYCA). The resulting raw sequencing data are provided in two FASTQ files, `OF.216F.i.raw\_1.fastq` and `OF.216F.i.raw\_2.fastq`, corresponding to the forward and reverse reads, respectively. This dataset offers a representative example for evaluating the performance and outputs of the PARAFISH platform. The studies presented in this paper were all conducted using Windows Subsystem for Linux (WSL) in a Linux environment (Ubuntu 22.04), which guaranteed complete compatibility with the external bioinformatics tools that were integrated into the PARAFISH platform. Most importantly, PARAFISH was created as a cross-platform solution, meaning that users can use the application on Windows, Linux, or MacOS as long as all necessary dependencies are loaded correctly and available through the system PATH. Preprocessing (including trimming and merging), sequence analysis (such as OTU clustering), taxonomic assignment using reference databases, phylogenetic inference, and visualization come after the initial inspection of raw sequencing data. The primary steps commonly seen in metabarcoding researches are included in this process, which is integrated into PARAFISH. The platform features an intuitive online interface and is designed with a modular architecture, allowing users flexibility in starting their analysis at various stages to suit their specific data and experimental requirements. Because of its adaptability and the ability to download intermediate and final results, PARAFISH can be used by a wide variety of users. By utilizing the platform interface to obtain complete logs of all operations and tool executions at every stage, users can additionally examine the analysis process and carry out any required troubleshooting. In the sections that follow, each step of the workflow will be fully described along with the results obtained from the selected dataset.

7.1 Quality Control

The PARAFISH workflow starts with the quality control assessment of raw sequencing data. The paired-end FASTQ files that match the sequencing reads of interest must be uploaded by users on the platform's "Quality Control" page. The platform makes it clear to users that when multiple files are provided, as in paired-end sequencing, FastQC analysis must be executed on each file separately. For both forward and reverse reads, this ensures that quality metrics are generated correctly. By selecting the appropriate file and clicking the "Run Quality Control" button after uploading it, the user initiates the quality control process. Each file receives an interactive HTML report from the platform's execution of FastQC, which can be seen and downloaded straight from the user interface. For the present study, the initial quality control of raw sequencing data for sample `OF.216F.i` was performed using FastQC. The forward reads (`OF.216F.i.raw_1.fastq.gz`) comprised 79,770 sequences with a length of 256 bp and a GC content of 46%. The reverse reads (`OF.216F.i.raw_2.fastq.gz`) also contained 79,770 sequences of 256 bp, with a GC content of 43%. No sequences were flagged as poor quality by FastQC's initial filter in either file. Analysis of the "Per base sequence quality" plots for both R1 and R2 reads revealed very high initial base call accuracy (median Phred scores generally above Q35) [35]. The quality did, however, noticeably decline around the 3'-ends. After position 210, the quality of the final 20 to 30 bases dropped into the warning area for R1 reads. The median stayed above Q28, but lower quartiles declined significantly. The quality reduction was more noticeable and began earlier for R2 reads, at position 190. The final 50–60 bases showed a substantial drop in quality, with median scores approaching Q32, but lower quartiles often dipped below Q28 and even Q20 at the very end. This pattern emphasizes the necessity of preprocessing 3'-end quality trimming. With reported adapter percentages of 0% across all locations for the tested adapter types, the "Adapter Content" module in FastQC was unable to identify any discernible presence of common Illumina or Nextera adapter sequences in either the R1 or R2 reads. Items like "Per base sequence content," "Sequence Duplication Levels," and "Overrepresented sequences" were recognized as warnings or failures by other FastQC modules. Given the targeted nature of PCR amplification along with potential primer effects at read ends, these flags are frequently seen in metabarcoding datasets [36]. They are usually addressed or comprehended in the context of downstream bioinformatics processing (e.g., primer trimming if not already done, and awareness of expected high-abundance amplicons) [37]. These FastQC results indicated that 3'-end quality trimming was the main preprocessing requirement in order to eliminate lower-quality bases and guarantee higher fidelity data for further analysis. A precautionary trimming step might still be taken in a standard process or if customized adapters were utilized, even when FastQC was unable to detect standard adapters.

7.2 Preprocessing

At the preprocessing stage, the PARAFISH platform offers users two distinct modes: single-end and paired-end. In single-end mode, the user uploads a single FASTQ file, which is processed exclusively with Trimmomatic for quality and adapter trimming. In paired-end mode, the user uploads two FASTQ files—one for each read direction—which are first processed with Trimmomatic and then merged using PEAR to generate high-quality consensus sequences. To ensure correct pairing, the platform requires that the uploaded files in paired-end mode are named with the suffixes `"_R1"` and `"_R2"` (for example, `OF.216F.i.raw_R1.fastq.gz` and `OF.216F.i.raw_R2.fastq.gz`), and users must rename their files accordingly before upload. The analyses described here were conducted in the paired-end manner to illustrate the entire workflow and its features. Users only need to choose the preprocessing option and start the analysis via the platform interface after uploading the paired-end FASTQ files with the correct names. The procedure uses a sliding window technique and conventional Illumina adapter sequences to apply Trimmomatic to eliminate low-quality bases and adapter contamination. For sample OF.216F.i, Trimmomatic processed 79,770 read pairs. Only roughly one percent of the reads were lost because they were too short or of poor quality after trimming, with approximately 96 percent of the pairings being successfully kept. This high retention rate reflects both the effectiveness of the trimming settings and the good initial quality of the data. PEAR then automatically retrieves the trimmed paired reads and combines the overlapping pairs into a single, high-quality consensus sequence. PEAR showed outstanding overlap and library preparation by merging more than 95 percent of the trimmed read pairs. The data's eligibility for further investigation is further supported by the extremely low percentage of unassembled or rejected reads. The flow of reads through these preprocessing steps, from raw input to merged consensus sequences, is visually summarized in Figure A1 (Annex A). Overall, the preprocessing results show that the PARAFISH platform optimizes data preservation and quality while effectively preparing sequencing data for further steps.

7.3 Sequence Analysis: OTU Clustering

Users can select from a variety of analytic techniques in the sequence analysis step using the PARAFISH platform. The VSEARCH OTU clustering approach was used for this investigation because of its effectiveness and widespread use in metabarcoding procedures. Users can choose anywhere from 80 to 100 for the sequence identity level for OTU clustering within the platform. Since 97 percent is frequently used to approximate species-level groups [38, 39], it was chosen as the default value for the current analysis. PARAFISH's flexibility is evident here: users can either upload their own processed FASTA file to initiate OTU clustering directly, or the platform will automatically utilize the FASTA-converted output from the preceding PEAR merging step. This adaptability, allowing users

to either leverage prior outputs or supply their own specific input files, is a consistent characteristic across all subsequent analytical modules. After choosing the VSEARCH OTU clustering method, users can start the process via the interface and adjust the identity threshold. Dereplication, which collapses identical sequences to decrease duplication [40,41], is the first process in the analysis. This dataset showed a significant reduction in redundancy and improved computing efficiency for the next clustering step by condensing 73,317 input sequences into 15,806 unique sequences. Clustering these unique sequences at the 97% identity threshold resulted in the identification of 63 OTUs. OTUs had a very unequal distribution of sequences; the largest OTU had 4,703 sequences, while 42 OTUs were singletons, each of which had a single sequence. This distribution, characterized by a few dominant OTUs and a long tail of rare ones, is clearly illustrated by the rank-abundance curve (Figure A2, Annex A) and a histogram of OTU sizes (Figure A3, Annex A). A few relatively frequent clusters with a lengthy tail of uncommon or singleton OTUs define this pattern, which is typical of environmental amplicon datasets and suggests the existence of rare taxa or sequencing artifacts in addition to biological diversity [42]. As a way to facilitate ecological analysis and subsequent taxonomic assignment, the platform automatically creates and makes available downloadable files that include the OTU representative sequences and the mapping of each read to its corresponding cluster.

7.4 Reference Databases: Taxonomic Assignment

In the reference database step, the PARAFISH platform allows users to select which database to use for the taxonomic assignment of their representative sequences. For this dataset, we performed taxonomic assignment using the BOLD database, which is particularly suitable for COI marker sequences due to its comprehensive coverage of animal barcodes [30]. The platform integrates BOLDigger, enabling users to customize their search by selecting several BOLD database options—such as the public animal library, species-level libraries, plant or fungi libraries, and others—according to the target taxa of their study. Users are also given the option to set custom identity thresholds for various taxonomic ranks (species, genus, family, order, and class); if these are left empty, BOLDigger's default thresholds are applied. Users must choose from three search modes: Rapid Species Search, Genus and Species Search, or Exhaustive Search. Once these settings are configured, the user starts the analysis, and the platform launches BOLDigger, delivering the findings as a Microsoft Excel document that can be downloaded. For the present analysis, the OTU sequences were compared against the BOLD animal library using the default thresholds and search mode. The majority of OTUs were allocated to the genus Trisopterus within the family Gadidae, according to the results, and many of them were successfully identified as Trisopterus luscus at the species level with high identity values (often over 97–98%). A small number of OTUs matched other taxa, such as Merluccius merluccius, while several OTUs could only be assigned at the genus or family level, and a few had no significant match in the database. No parasites were detected

in the sample, but several fish were found. The overall distribution of these taxonomic assignment levels is presented in Figure A4 (Annex A). This pattern is typical for environmental COI datasets, reflecting both the presence of dominant taxa and the limitations of reference coverage for rare or novel sequences [43]. The platform's output table includes, for each OTU, the assigned taxonomy, percent identity, number of records supporting the match, and the BOLD BIN identifier when available, facilitating downstream ecological interpretation. The distribution of the percent identity scores for these assignments, which generally indicates the quality of the matches, is visualized in Figure A5 (Annex A).

7.5 Phylogenetic Analysis

The PARAFISH platform automates a robust workflow in the phylogenetic analysis process, including MAFFT for multiple sequence alignment, FastTree2 for fast tree inference and RAxML-NG for maximum likelihood phylogenetic reconstruction. MAFFT aligns the input sequences using the L-INS-i approach, which is well-known for its high accuracy [44]. All alignment methods, including scoring, guide tree construction, and iterative refinement, were successfully completed by MAFFT for the current dataset of 63 COI sequences. Using the GTR+CAT model, which is suitable for nucleotide data like COI, FastTree2 quickly infers an initial phylogenetic tree following alignment. FastTree swiftly optimized the tree topology, ensuring a robust basis for analysis. RAxML-NG allows users to adjust the number of bootstrap replicates, enabling a trade-off between computational time and the robustness of statistical support for tree branches [45–47]. To guarantee robust statistical support for phylogenetic relationships, 500 bootstrap replicates were selected. This number is frequently advised for datasets of moderate size and offers a fair balance between efficiency and accuracy [46]. Upon completing, the platform creates and makes the outputs from each tool accessible for download: the original tree from FastTree2, visualized in Figure A6 (Annex A), the best-scoring maximum likelihood tree from RAxML-NG, and the aligned sequences from MAFFT. Users can also view both phylogenetic trees directly on the platform through using a breadthfirst arrangement, which arranges the tree in layers from the root outward, making it simple to understand the branching patterns and hierarchical relationships [48]. Cytoscape.js was used to create this interactive display, which improves the results' interpretability by enabling users to quickly and easily explore the tree structure. The final results attest to the satisfactory completion of every stage of the workflow and the suitability of the selected techniques for COI metabarcoding data. Some branches in the phylogenetic tree generated by RAxML-NG exhibited nearly zero lengths, suggesting highly similar or unresolved sequences [49].

8 Platform Performance and Comparison

This section evaluates the platform's computational performance and presents a comparison with PMiFish after a thorough demonstration of the PARAFISH

workflow using the `OF.216F.i` dataset. The purpose of this analysis is to provide information about PARAFISH's unique contributions to the field as well as its operational effectiveness.

8.1 Performance Benchmarks

Benchmarking Methodology
The previously described `OF.216F.i` sample dataset was used to benchmark the computational performance of each major step in the PARAFISH workflow. The Unix `time` command for external tool invocation and Python's internal time module were utilized to document execution times. System utilities were used for monitoring peak CPU and RAM utilization. All benchmarks were run on a system with an AMD Ryzen 5 5600H CPU (6 cores, 12 threads) and approximately 2.8 GB RAM allocated to the WSL2 Ubuntu 22.04 environment.

Benchmark Results and Analysis
The detailed performance metrics for processing the `OF.216F.i` sample are presented in Table A1 (Annex A). Performance metrics (Table A1) indicate that PARAFISH efficiently processes a typical COI metabarcoding dataset. The PARAFISH platform demonstrated efficient computational performance across various stages. Key tools, including Trimmomatic, FastQC, and VSEARCH, optimized CPU usage and memory within 2.8 GB limits. The most time-intensive steps were PEAR for merging paired-end reads (7 minutes) and RAxML-NG for maximum likelihood tree inference with 500 bootstraps (44.52 seconds). The workflow exhibits suitability for typical lab PCs handling datasets of this size.

8.2 Comparison with Other Tools: PMiFish

To position PARAFISH within the existing bioinformatics landscape, its features were compared against PMiFish. A summary of this comparison is provided in Table A2 (Annex A). Both PMiFish and PARAFISH offer complete solutions for analyzing metabarcoding data, covering everything from taxonomic identification to quality control. PARAFISH is a web-based, user-friendly platform for metabarcoding analysis, designed to simplify biodiversity studies, particularly in fish parasite research. It integrates tools like MAFFT, FastTree2, and RAxML-NG for comprehensive phylogenetic analysis, offering interactive tree visualizations. Unlike PMiFish, which is command-line based and optimized for MiFish markers [13], PARAFISH provides a graphical interface and modular workflow, allowing flexibility and accessibility. It balances computational efficiency with robust analytical capabilities and promises future adaptability for expanded research applications, supporting ecological studies, aquaculture, and food safety regulation.

References

1. Timi, J.T., Poulin, R.: Why ignoring parasites in fish ecology is a mistake. *International Journal for Parasitology* **50**(10–11), 755–761 (2020). doi:10.1016/j.ijpara.2020.04.007.
2. Gold, Z., Curd, E.E., Goodwin, K.D., et al.: Improving metabarcoding taxonomic assignment: a case study of fishes in a large marine ecosystem. *Molecular Ecology Resources* **21**(7), 2546–2564 (2021). doi:10.1111/1755-0998.13450.
3. Aivelo, T., Medlar, A.: Opportunities and challenges in metabarcoding approaches for helminth community identification in wild mammals. *Parasitology* **145**(5), 608–621 (2018). doi:10.1017/S0031182017000610.
4. Miller, M.L., Rota, C., Welsh, A.: Transforming gastrointestinal helminth parasite identification in vertebrate hosts with metabarcoding: a systematic review. *Parasites & Vectors* **17**(1) (2024). doi:10.1186/S13071-024-06388-1.
5. Taberlet, P., Coissac, E., Pompanon, F., et al.: Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology* **21**(8), 2045–2050 (2012). doi:10.1111/j.1365-294X.2012.05470.x.
6. Coissac, E., Riaz, T., Puillandre, N.: Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology* **21**(8), 1834–1847 (2012). doi:10.1111/j.1365-294X.2012.05550.x.
7. Chan, A.H.E., Saralamba, N., Saralamba, S., et al.: Sensitive and accurate DNA metabarcoding of parasitic helminth mock communities using the mitochondrial rRNA genes. *Scientific Reports* **12**(1) (2022). doi:10.1038/s41598-022-14176-z.
8. Surmacz, B., Vecchi, M., Fontaneto, D., et al.: COI metabarcoding with a curated reference database and optimized protocol provides a reliable species-level diversity assessment of tardigrades. *Integrative Zoology* (2025). doi:10.1111/1749-4877.12972.
9. Bourret, A., Nozères, C., Parent, E., et al.: Maximizing the reliability and the number of species assignments in metabarcoding studies using a curated regional library and a public repository. *Metabarcoding and Metagenomics* **7**, 37–49 (2023). doi:10.3897/mbmg.7.98539.
10. Fontes, J.T., Vieira, P.E., Ekrem, T., et al.: BAGS: an automated barcode, audit & grade system for DNA barcode reference libraries. *Molecular Ecology Resources* **21**(2), 573–583 (2021). doi:10.1111/1755-0998.13262.
11. Whitfield, S., Hofmann, M.A.: Elicit: AI literature review research assistant. *Public Services Quarterly* **19**(3), 201–207 (2023). doi:10.1080/15228959.2023.2224125.
12. Claver, C., Canals, O., de Amézaga, L.G., et al.: An automated workflow to assess completeness and curate GenBank for eDNA metabarcoding: the marine fish assemblage as case study. *bioRxiv* (2022). doi:10.1101/2022.10.26.513819.
13. rogotoh: PMiFish. https://github.com/rogotoh/PMiFish (Accessed: June 20, 2025).
14. Brown, J., Pirrung, M., McCue, L.A.: FQC dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**(19), 3137–3139 (2017). doi:10.1093/bioinformatics/btx373.
15. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**(15), 2114–2120 (2014). doi:10.1093/bioinformatics/btu170.
16. Zhang, J., Kobert, K., Flouri, T., et al.: PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**(5), 614–620 (2014). doi:10.1093/bioinformatics/btt593.

17. de Santiago, A., Pereira, T.J., Mincks, S.L., et al.: Dataset complexity impacts both MOTU delimitation and biodiversity estimates in eukaryotic 18S rRNA metabarcoding studies. *Environmental DNA* **4**(2), 363–384 (2022). doi:10.1002/edn3.255.
18. Chiarello, M., McCauley, M., Villéger, S., et al.: 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 (2022). doi:10.1371/journal.pone.0264443.
19. Prodan, A., Tremaroli, V., Brolin, H., et al.: Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PloS One* **15**(1), e0227434 (2020). doi:10.1371/journal.pone.0227434.
20. Bhat, A.H., Prabhu, P.: OTU clustering: a window to analyse uncultured microbial world. *International Journal of Scientific Research in Computer Science and Engineering* **5**(6), 62–68 (2017). doi:10.26438/ijsrcse/v5i6.6268.
21. Rognes, T., Flouri, T., Nichols, B., et al.: VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016). doi:10.7717/peerj.2584.
22. Nearing, J.T., Douglas, G.M., Comeau, A.M., et al.: Denoising the denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**, e5364 (2018). doi:10.7717/peerj.5364.
23. Callahan, B.J., McMurdie, P.J., Rosen, M.J., et al.: DADA2: high resolution sample inference from Illumina amplicon data. *Nature Methods* **13**(7), 581–583 (2016). doi:10.1038/nmeth.3869.
24. Jurado-Rueda, F., Alonso-Guirado, L., Perea-Chamblee, T.E., et al.: Benchmarking of microbiome detection tools on RNA-seq synthetic databases according to diverse conditions. *Bioinformatics Advances* **3**(1), vbad014 (2023). doi:10.1093/bioadv/vbad014.
25. Straub, D., Blackwell, N., Langarica-Fuentes, A., et al.: Interpretations of environmental microbial community studies are biased by the selected 16S rRNA (gene) amplicon sequencing pipeline. *Frontiers in Microbiology* **11**, 550420 (2020). doi:10.3389/fmicb.2020.550420.
26. Lu, J., Salzberg, S.L.: Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome* **8**(1), 900 (2020). doi:10.1186/s40168-020-00900-2.
27. Vaulot, D., Sim, C.W.H., Ong, D., et al.: metaPR2: a database of eukaryotic 18S rRNA metabarcodes with an emphasis on protists. *Molecular Ecology Resources* **22**(8), 3188–3201 (2022). doi:10.1111/1755-0998.13674.
28. Mugnai, F., Costantini, F., Chenuil, A., et al.: Be positive: customized reference databases and new, local barcodes balance false taxonomic assignments in metabarcoding studies. *PeerJ* **11**, e14616 (2023). doi:10.7717/peerj.14616.
29. Jeunen, G.J., Dowle, E., Edgecombe, J., et al.: Crabs-A software program to generate curated reference databases for metabarcoding sequencing data. *Molecular Ecology Resources* **23**(3), 725–738 (2023). doi:10.1111/1755-0998.13741.
30. Meglécz, E.: COInr and mkCOInr: building and customizing a nonredundant barcoding reference database from BOLD and NCBI using a semi-automated pipeline. *Molecular Ecology Resources* **23**(4), 933–945 (2023). doi:10.1111/1755-0998.13756.
31. O'Rourke, D.R., Bokulich, N.A., Jusino, M.A., et al.: A total crapshoot? evaluating bioinformatic decisions in animal diet metabarcoding analyses. *Ecology and Evolution* **10**(18), 9721–9739 (2020). doi:10.1002/ece3.6594.
32. Buchner, D., Leese, F.: BOLDigger – a Python package to identify and organise sequences with the Barcode of Life Data systems. *Metabarcoding and Metagenomics* **4**, e53535 (2020). doi:10.3897/mbmg.4.53535.

33. Young, C., Meng, S., Moshiri, N.: An evaluation of phylogenetic workflows in viral molecular epidemiology. *Viruses* **14**(4), 774 (2022). doi:10.3390/v14040774.
34. Hossain, S.: Visualization of bioinformatics data with Dash Bio. *Proceedings of the 18th Python in Science Conference*, 126–133 (2019). doi:10.25080/majora-7ddc1dd1-012.
35. Ewing, B., Green, P.: Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Research* **8**(3), 186–194 (1998). doi:10.1101/gr.8.3.186.
36. Elbrecht, V., Hebert, P.D.N., Steinke, D.: Slippage of degenerate primers can cause variation in amplicon length. *Scientific Reports* **8**(1), 29364 (2018). doi:10.1038/s41598-018-29364-z.
37. Kechin, A., Boyarskikh, U., Kel, A., et al.: CutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing. *Journal of Computational Biology* **24**(11), 1138–1143 (2017). doi:10.1089/cmb.2017.0096.
38. Edgar, R.C.: Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* **34**(14), 2371–2375 (2018). doi:10.1093/bioinformatics/bty113.
39. Li, X., Huo, S., Xi, B.: Updating the resolution for 16S rRNA OTUs clustering reveals the cryptic cyanobacterial genus and species. *Ecological Indicators* **117**, 106695 (2020). doi:10.1016/j.ecolind.2020.106695.
40. Seguritan, V., Rohwer, F.: FastGroup: a program to dereplicate libraries of 16S rDNA sequences. *BMC Bioinformatics* **2**(1), 9 (2001). doi:10.1186/1471-2105-2-9.
41. Burriesci, M.S., Lehnert, E.M., Pringle, J.R.: Fulcrum: condensing redundant reads from high-throughput sequencing studies. *Bioinformatics* **28**(10), 1324–1327 (2012). doi:10.1093/bioinformatics/bts123.
42. Huse, S.M., Welch, D.M., Morrison, H.G., et al.: Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* **12**(7), 1889–1898 (2010). doi:10.1111/j.1462-2920.2010.02193.x.
43. Zafeiropoulos, H., Gargan, L., Hintikka, S., et al.: The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data. *Metabarcoding and Metagenomics* **5**, e69657 (2021). doi:10.3897/mbmg.5.69657.
44. Katoh, K., Kuma, K., Miyata, T., et al.: Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Informatics* **16**(1), 22–33 (2005). doi:10.11234/gi1990.16.22.
45. Lemoine, F., Domelevo Entfellner, J.B., Wilkinson, E., et al.: Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**(7702), 452–456 (2018). doi:10.1038/s41586-018-0043-0.
46. Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., et al.: How many bootstrap replicates are necessary? *Lecture Notes in Computer Science* **5541**, 184–200 (2009). doi:10.1007/978-3-642-02008-$7_1$3.
47. Stamatakis, A., Hoover, P., Rougemont, J.: A rapid bootstrap algorithm for the RAxML web servers. *Systematic Biology* **57**(5), 758–771 (2008). doi:10.1080/10635150802429642.
48. Andrusky, K., Curial, S., Amaral, J.N.: Tree-traversal orientation analysis. *Lecture Notes in Computer Science* **4382**, 220–234 (2007). doi:10.1007/978-3-540-72521-$3_1$7.
49. Kozlov, A.M., Darriba, D., Flouri, T., et al.: RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**(21), 4453–4455 (2019). doi:10.1093/bioinformatics/btz305.

A Supplementary Information

The supplementary materials listed below are available in the project's GitHub repository.

A.1 Supplementary Figures

- **Figure A1:** Read retention flowchart through preprocessing steps (Trimmomatic and PEAR) for sample OF.216F.i.
  File: `annexes/figures/sankey_preprocessing.png`
- **Figure A2:** OTU rank-abundance curve for the 63 OTUs identified from sample OF.216F.i, illustrating the dominance of a few OTUs and a long tail of rare OTUs.
  File: `annexes/figures/figure_otu_rank_abundance.png`
- **Figure A3:** Histogram of OTU sizes for sample OF.216F.i, showing the frequency distribution of OTUs based on the number of sequences they contain.
  File: `annexes/figures/figure_otu_histogram.png`
- **Figure A4:** Distribution of OTU Taxonomic Assignment Levels from BOLDigger.
  File: `annexes/figures/assignment_levels.png`
- **Figure A5:** Distribution of Percent Identity Scores for Taxonomic Assignments via BOLDigger.
  File: `annexes/figures/pct_identity_histogram.png`
- **Figure A6:** Example of the interactive phylogenetic tree visualization within the PARAFISH platform for the 63 OTUs.
  File: `annexes/figures/figure_parafish_tree_visualization.png`

A.2 Supplementary Tables

- **Table A1:** Performance Metrics for PARAFISH Workflow using Sample OF.216F.i.
  File: `annexes/figures/Table_A1.pdf`
- **Table A2:** Feature Comparison between PARAFISH and PMiFish.
  File: `annexes/figures/Table_A2.pdf`