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. As the capacity to sequence tens to hundreds of millions of reads in a single sequencing run enables the creation of new research questions, such as species mapping, biomonitoring, gut content analyses, and population genomics, metabarcoding is being utilized increasingly in the study 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 Taberlet [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 most cases, the reference data collection is not supervised or checked for quality control, with a few exceptions. Erroneous data in reference libraries can cause repeated and undetected identification errors over time [8]. To ensure our review reflects the latest methodologies and technologies, we employed Elicit AI, an LLM-based literature review search engine to support the research process [9], to identify recent advances in metabarcoding and reference database curation. 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. Manual database curation becomes difficult in marine habitats due to their size and significant diversity, in contrast to the limited and well-documented freshwater ecosystems. In recent years, metabarcoding of eDNA has become a more common technique for studying marine fish, as it has shown great potential as an alternative monitoring tool. In this regard, using eDNA-based techniques in fisheries monitoring would depend heavily on the availability of comprehensive and maintained datasets. Universal metazoan 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 [10].

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. 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. This project aims to use advanced bioinformatics software to process large datasets with precision and speed. Technologies such as next-generation sequencing (NGS) will allow us to obtain detailed genetic information about parasites, which will then be analyzed using algorithms that can detect patterns and anomalies.

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 [11]. Then we will perform trimming with Trimmomatic to remove low-quality regions and adapter sequences [12]. 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) [13]. 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 OTUs and ASVs in their entirety and evaluating their influence on bioinformatics processes [14, 15]. 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 [14]. One example of such an algorithm is USEARCH [16, 17], 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 [18]. 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 amplicon sequence variants (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 [14], and when compared to Deblur, DADA2 typically generates more Amplicon Sequence Variants (ASVs) [19]. DADA2, an open-source R package, transforms and enhances the DADA algorithm [20] 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 [16]. 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 [21,22]. Additionally, Kraken2 supports 16S rRNA classification using databases such as Greengenes, SILVA, and RDP [23]. 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 [24]. Additionally, the cytochrome oxidase subunit I (COI) gene and other ribosomal RNA genes (16S and 12S) are also essential markers for eukaryotic research [25]. 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 [24, 26]. 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 Barcode of Life 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 Barcode of Life Data System (BOLD), which primarily holds the COI gene's barcoding segment even though it also contains sequences for other markers [27]. The COI animal marker reference datasets from GenBank and BOLD are complementary to one another; however, since both databases are updated continuously without a fixed version history, unlike microbial reference databases, studies carried out within a short period may report differences of thousands of reference sequences [28]. 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 [10]. 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 [29]. 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 [30]. This workflow will produce phylogenetic trees, illustrating evolutionary relationships among identified taxa. 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 [31], 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.

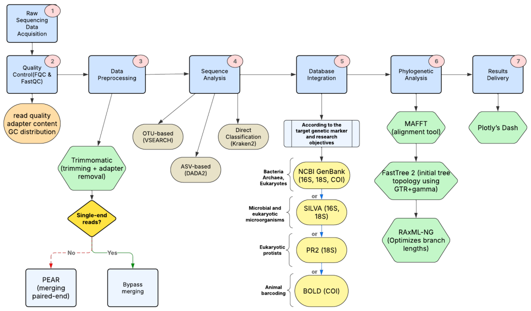
Raw Sequencing Data Acquisition
1
Quality Control(FQC & FastQC)
2
read quality adapter content GC distribution
Data Preprocessing
3
Trimmomatic (trimming + adapter removal)
Single-end reads?
No
Yes
PEAR (merging paired-end)
Bypass merging
Sequence Analysis
4
OTU-based (VSEARCH)
ASV-based (DADA2)
Direct Classification (Kraken2)
Database Integration
5
According to the target genetic marker and research objectives
Bacteria Archaea, Eukaryotes
NCBI GenBank (16S, 18S, COI)
or
Microbial and eukaryotic microorganisms
SILVA (16S, 18S)
or
Eukaryotic protists
PR2 (18S)
or
Animal barcoding
BOLD (COI)
Phylogenetic Analysis
6
MAFFT (alignment tool)
FastTree 2 (initial tree topology using GTR+gamma)
RAxML-NG (Optimizes branch lengths)
Results Delivery
7
Plotly's Dash

Fig. 1. Workflow diagram of the metabarcoding platform.

6 Expected Outcomes

As a result, 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 through long-term integration will benefit aquaculture, researchers, and food safety regulators.

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., Brochmann, C., Willerslev, E.: 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. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. Zhang, J., Kobert, K., Flouri, T., Stamatakis, A.: PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**(5), 614–620 (2014). doi:10.1093/bioinformatics/btt593.
14. De Santiago, A., Pereira, T.J., Mincks, S.L., Bik, H.M.: 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.
15. Chiarello, M., McCauley, M., Villéger, S., Jackson, C.R.: 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) (2022). doi:10.1371/journal.pone.0264443.
16. Prodan, A., Tremaroli, V., Brolin, H., et al.: Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PloS One* **15**(1) (2020). doi:10.1371/journal.pone.0227434.

17. 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.
18. Rognes, T., Flouri, T., Nichols, B., Quince, C., Mahé, F.: VSEARCH: A versatile open source tool for metagenomics. *PeerJ* **2016**(10), e2584 (2016). doi:10.7717/peerj.2584.
19. Nearing, J.T., Douglas, G.M., Comeau, A.M., Langille, M.G.I.: Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**(8), e5364 (2018). doi:10.7717/peerj.5364.
20. 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 (2016). doi:10.1038/nmeth.3869.
21. Jurado-Rueda, F., Alonso-Guirado, L., Perea-Chamblee, T.E., Elliott, O.T., Filip, I., Rabadá, R., Malats, N.: Benchmarking of microbiome detection tools on RNA-seq synthetic databases according to diverse conditions. *Bioinformatics Advances* (2023). doi:10.1093/bioadv/vbad014.
22. Straub, D., Blackwell, N., Langarica-Fuentes, A., Peltzer, A., Nahnsen, S., Kleindienst, S.: Interpretations of environmental microbial community studies are biased by the selected 16S rRNA (gene) amplicon sequencing pipeline. *Frontiers in Microbiology* **11** (2020). doi:10.3389/fmicb.2020.550420.
23. Lu, J., Salzberg, S.L.: Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. *Microbiome* **8**(1) (2020). doi:10.1186/S40168-020-00900-2.
24. Vaulot, D., Sim, C.W.H., Ong, D., Teo, B., Biwer, C., Jamy, M., Lopes dos Santos, A.: 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.
25. Mugnai, F., Costantini, F., Chenuil, A., Leduc, M., Gutiérrez Ortega, J.M., Meglécz, E.: 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/SUPP-15.
26. Jeunen, G.J., Dowle, E., Edgecombe, J., von Ammon, U., Gemmell, N.J., Cross, H.: 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.
27. 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.
28. O'Rourke, D.R., Bokulich, N.A., Jusino, M.A., MacManes, M.D., Foster, J.T.: A total crapshoot? Evaluating bioinformatic decisions in animal diet metabarcoding analyses. *Ecology and Evolution* **10**(18), 9721–9739 (2020). doi:10.1002/ECE3.6594.
29. 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.
30. Young, C., Meng, S., Moshiri, N.: An evaluation of phylogenetic workflows in viral molecular epidemiology. *Viruses* **14**(4) (2022). doi:10.3390/V14040774.
31. 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.