Customization of bioinformatics workflows for (meta)genomics

Ricardo Oliveira[1] and Pedro Santos[2]

[1] School of Engineering, Minho University, Campus de Azurém, 4804 - 533 Guimarães, Portugal
[2] Biology Department, Minho University, Campus de Gualtar, 4710-057 Braga, Portugal

Abstract. Metabarcoding has revolutionized biodiversity analysis by enabling rapid taxonomic classification from complex environmental samples. However, the proliferation of software tools and algorithms often leads to workflow inconsistencies and potential biases. This study addresses this challenge by evaluating various processing steps and tools to identify optimal solutions for a robust and automated metabarcoding analysis pipeline. We developed an R script to assess the performance of different filtration parameters, taxonomic assignment tools, and reference databases using mock communities. High-performing options were then integrated into a user-friendly, modular pipeline using R and Python with four key components: user input, pre-processing and amplicon sequence variant formation, taxonomic assignment, and results exploration. The new pipeline demonstrates accurate analysis for mock communities Zymo and ATCC, highlighting its potential for real-world applications. However, further validation with diverse environmental samples is necessary to ensure generalizability. This work paves the way for standardized and user-friendly metabarcoding analysis, promoting reliable taxonomic classification in ecological studies.

Keywords: Metabarcoding, taxonomic classification, automated pipeline

1 Introduction

1.1 Background

Identifying and characterizing microorganisms are fundamental aspects of microbiology, essential for perceiving how microorganisms produce or transform compounds of interest, for biosurveillance, assessing the quality of industrial processes, or discovering new strains with biotechnological applications [1, 2]. The classic approach identified microorganisms based on their physical and cultural traits, morphology, physiology, and biochemical characteristics. However, the advances in computational power, molecular biology, and sequencing technologies have led to the development of a revolutionary DNA-based approach unlocking metagenomics on a large scale with improved efficiency, reproducibility, and the ability to identify and characterize strains of microorganisms that are difficult to culture or may not grow in a laboratory environment [3].

Recent developments in DNA-based approaches, more specifically in high-throughput sequencing technologies, have notably increased the application of molecular methods, such as metabarcoding (identification of the different taxa present in complex samples using marker genes), resulting in a massive increase of the amounts of biological data produced [4]. In response, a vast diversity of bioinformatics tools have been designed to allow the analysis of metabarcoding data. QIIME1, MOTHUR and USEARCH software packages were among the pioneering solutions to interpret metabarcoding data, specially for bacterial 16S rRNA amplicon analysis. Nevertheless, the applications of metabarcoding have been expanded to a wide diversity of taxa from the most diverse environments, resulting in a significant increase in pipelines [4]. Consequently, two major groups of pipelines were formed. The first group comprises pipelines based on software that provides a highly customizable set of algorithms, such as DADA2 [5], and USEARCH/VSEARCH [6] while, the second is a set of pipelines, such as PIMBA [7] which provides a pre-defined set of algorithms previously validated on specific sequencing data to facilitate metabarcoding data analysis for users with limited bioinformatics knowledge. Nonetheless, these tools can deliver a wrong analysis when the dataset analyzed differs from the one used to validate the pipelines [4].

The rapid increase in bioinformatics tools and pipelines, while crucial for data analysis, also presents one of the bottlenecks when performing metabarcoding analysis, as it adds complexity in choosing how to process the datasets in relevant features [4]. In addition, the most common method of microbial identification is based on the amplification and sequencing of target genes. Consequently, most of the tools developed were designed to analyze amplicon data sequencing and are unsuitable for data generated from other approaches, such as whole-shotgun [8]. Overall, a standard workflow for amplicon data is followed to perform metabarcoding analysis varying the algorithm choice depending on the dataset's characteristics [4]. Each phase of this type of workflow will be analyzed below, along with the problems and biases faced and how the different algorithms overcome them.

1.2 Quality control and pre-processing

The initial stage of metabarcoding data processing incorporates quality control and pre-processing of raw data, which is a vital procedure to ensure the accuracy of the analysis. This phase involves various parameters that considerably impact the output. Upon demultiplexing, sequencing adapters and other non-biological information such as primers must be removed with tools like Trimmomatic or AdapterRemoval to eliminate nucleotides added during the amplification and sequencing, as they do not provide useful information and can lead to wrong conclusions [4].

Significantly, during a metabarcoding workflow, at least three sources of error can be identified: DNA degradation, PCR-generated errors, and sequencing errors. All these errors add noise to the data and the algorithms used for error removal, although essential for an accurate analysis, are often biased [10]. To

remove sequencing errors, softwares, such as DADA2, incorporate a sequence filtration algorithm, usually based on the read quality score. However, this is coarse filtering where parameters such as read length, number of ambiguous bases, or average quality score should be carefully considered, especially for datasets with higher sequencing errors. Excessively conservative parameters can result in the removal of too many reads (or bases) and therefore loss of sensitivity to low-abundance taxa [11]. The sequence filtration can be done based on per-sequence quality, discarding the whole sequence if it does not meet the threshold, or based on per-nucleotide quality truncating the sequence from the position below the threshold keeping a partial amplicon. The threshold requirement is usually based on the sum of the error probabilities because a good average quality score can mask several bases with high error probabilities which can be interpreted as biological variations [4].

Following sequence filtration, many pipelines include two clustering algorithms to reduce PCR amplification errors, since the amplification can introduce single-base substitutions and length variations that cannot be associated with low-quality reads and can be misinterpreted as biological variation substantially increasing downstream computation [8, 11]. Nonetheless, these denoising algorithms tend to discard low-abundance sequences which are likely to be artifacts, which, in some scenarios, is hard to distinguish noise from a real low-abundance signal therefore discarding rare taxa, especially for datasets with low sequencing depth [4]. Nowadays, there is an open discussion in the scientific community to determine which algorithm should be used to perform error removal without bias but there is no unanimous decision. However, the use of negative controls can be useful to eliminate contaminants introduced during the experimental method. Nevertheless, there is also no consensus on how to handle the negative controls bioinformatically and there is a debate on when contamination is found the entire dataset should be eliminated. In addition, mock community samples are a way of determining experimental bias [4, 12].

1.3 Data clustering

The operational taxonomic units (OTU) are a clustering approach that clusters all the sequences within an arbitrary similarity threshold commonly fixed to 97 percent. Nonetheless, a 97 percent similarity threshold is thought to be a balanced compromise between interspecific and intraspecific variation, when a single threshold is applied it can split a single species across various OTUs or lump multiple species into the same OTU [4, 11]. Standard workflows depended on clustering using OTUs, however, this approach does not take into consideration individual sequence quality information, and therefore, using an arbitrary threshold can add bias to the results since error rates often vary substantially between sequencing runs and PCR protocols [8]. To solve this problem, new denoising algorithms that integrate statistical models and sequence quality information were developed to distinguish true biological sequences from sequencing noise. The clustering of amplicon sequence variants (ASV) is a new denoising

algorithm present in DADA2, which relies on a parameterized model of substitution errors to distinguish sequencing errors from real biological variation joining together reads with as few as 1 base pair difference between variants representing biological sequences before amplification and sequencing errors [8]. Furthermore, denoising methods are not designed to deal with high-abundance artifacts like PCR-generated chimeras and non-specific amplification products and need supplementary algorithms. Chimeras can be algorithmically removed by comparison with the other reads (de novo) or with a chimera-free reference database.

1.4 Taxonomic assignment

After assuring that the data meets the quality standards, the next stage of the workflow corresponds to the assignment of Linnaean taxonomy to each sequence. This process is automated by a wide variety of software based on two major approaches. The most widely used approach for taxonomic classification is the best-hit classification resorting to alignment-based tools such as BLAST. This approach assumes that the taxonomy of the query sequence will be identical to the taxonomy of the most similar sequence in a reference database. Although this approach is simple to implement and effective when reference data is unavailable best-hit classification is likely to over-classify the sequence to an incorrect species-level taxonomy leading to the identification of unsequenced organisms [11]. In alternative, algorithms, such as RDP Naïve Bayesian classifier, can be used to realize the taxonomic classification based on the sequence composition where a set of features, such as GC content or oligonucleotide sequence, is compared to a reference database of genomes [13]. Nonetheless, composition-based approaches tend to be more prone to errors for less complete databases than alignment-based methods, although for higher taxonomic ranks, like family level, both methods have similar performance [4]. To improve the assignment accuracy some tools implement a hybrid approach utilizing global alignment tools and SINTAX to estimate the consensus last common ancestor taxonomy [4].

Various studies tested the accuracy of different taxonomy assignment methods and reported that the major cause of errors during this phase is incomplete reference databases establishing a relationship between the assignment accuracy and database completeness [4]. Databases such as ribosomal database project, SILVA, and Greengenes [14] are the most widely used sequence databases for DNA metabarcoding analysis. However, the coverage of those databases is incomplete and uneven with extreme variation based on the taxonomic group [15]. For instance, studies by López-Escardó et al (2018) and Bakker et al. (2019) stated that database coverage for marine species is less complete, with only 30 percent of the OTUs having matches at the species level and only half of the reads were able to be identified to the phylum level. Moreover, regardless of the assignment method used is important to select a database that also includes nontarget taxa to limit overclassification [4].

In order to evaluate the accuracy of the taxonomic assignment and optimize the pipelines benchmark studies must be done. For such purpose, mock commu-

nities containing a known heterogeneous composition including varying numbers of taxonomic groups and individuals per taxa should be used to mimic the natural ecosystem [12]. The accuracy of the taxonomic assignment can be measured through a defined set of metrics, such as scoring false discovery rates, true positive rates, and Matthews correlation coefficient. The results of the benchmark studies can be used to evaluate the confidence in the metabarcoding analysis and, if the results do not meet defined quality standards, they can be used to test modifications, such as different pre-processing algorithms and sequence filtering approaches, to optimize the pipelines for an accurate and precise metabarcoding analysis [17]. Although mock communities are necessary to evaluate the bioinformatics performance and the experimental protocols, they also can impose bias if not well-chosen. For instance, since most commercial mocks are mainly focused on the human microbiome they may not have a relevant taxonomic representation for the environment of interest. Thus, customized mocks aimed at the environment under study must be prepared in the laboratory [12].

1.5 Data exploitation

Effective data visualization through tools such as Phyloseq is a crucial phase for interpreting metabarcoding results since it facilitates the interpretation of complex relationships among microbial communities. In addition, the comprehensive exploration of the results can also be used to detect biases introduced in the previous phases and to identify samples and microorganisms that might need special attention and exploration, or, in some cases, a re-analysis [8]. Diversity indices are central components of exploiting metabarcoding data allowing the quantification and comparison of the biodiversity observed in different environments through the employment of various metrics, such as alpha diversity or beta diversity. Alpha diversity uses a set of metrics that include the Shannon index, Simpson index, and observed richness to provide valuable insights into the complexity and evenness of microbial communities within each sample revealing important variations in microbial diversity associated with different environmental conditions [8]. On the other hand, beta diversity is used to identify ecological patterns and relationships among microbial communities through ordination methods such as principal coordinates analysis (PCoA), and non-metric multidimensional scaling (NMDS) [8]. Taxonomic composition analysis allows the understanding of microbial community structure via visualization of relative abundances of different taxa at various taxonomic levels from phylum to species, highlighting the dominant microbial groups under the environment conditions under study [8]. Differential analysis is used to identify the taxa that are differentially abundant between groups of samples, being a crucial approach to detecting significant changes in the abundance of specific taxa, providing precious knowledge into microbial responses to different environmental conditions [8].

1.6 Objective

Our project aimed to evaluate the performance of metabarcoding analysis tools using datasets generated with different sequencing platforms (Illumina, PacBio, and Oxford Nanopore) systematically. We have assessed each step of the analysis workflow using DADA2 [5] to determine the most accurate parameterization for every step of the workflow. For this purpose, we resorted to commercial mock community sequencing datasets (e.g. Zymobiomics and ATCC). Based on the generated information, we have developed a set of customizable scripts that use the best tool for each phase of the analysis process by default. Nevertheless, the user will have the option to choose which tool to use and set the desired parameters for every analysis phase. Furthermore, our scripts will incorporate a set of metrics that allow measure the performance of the chosen workflow and evaluate if the results meet the established standards or if changes to the workflow are necessary to provide a more accurate result. Finally, we plan to apply the developed scripts to a real case study in our laboratory.

2 Materials and Methods

2.1 Datasets

Zymo Research's D6300 and D6305 mock communities (Zymo1 and Zymo2), were utilized for performance analysis and benchmark studies since these communities are established standards in metabarcoding analysis, resulting in wide availability of data produced by Illumina, PacBio, and nanopore sequencing. These communities comprise ten organisms with abundances varying from 12 percent to 2 percent, allowing the assessment of the pipeline's capability to identify organisms with low abundances. ATCC (ATCC1, ATCC2, and ATCC3) mock communities were also used, to evaluate the sensibility of the developed pipeline to discriminate different microorganisms since these communities are constituted by 20 species of bacteria with even abundance. All datasets were retrieved from the National Center for Biotechnology Information (NCBI) and were generated by sequencing the V4 region of the 16S rRNA gene, with the Illumina platform.

2.2 Personalized Scripts

In order to develop an automated method for evaluating the performance of different amplicon analysis software and to create a user-friendly tool that doesn't require extensive bioinformatics expertise, allowing laboratory staff to conduct a default amplicon analysis and utilize automation, a set of scripts using Python (version 3.12.4) and R (version 4.3.3) were created. These scripts were designed with a modular architecture, encompassing four distinct modules, allowing users to choose which analysis phase to perform within the amplicon analysis workflow.

User Inputs: The first module was written in Python and was designed to prompt the user to provide the necessary information to proceed with the analysis. Initially, the user is prompted to select an action among pre-processing the data and creation of ASVs, performing taxonomic assignment of the ASVs, or exploring the obtained results. Once the user specifies the intended action, a new set of prompts gathers information about the directories where the files to analyze are located, file naming, and saving preferences. For pre-processing and ASV formation, details about the library construction, such as the sequencing technology and the specific region of the 16S rRNA that was sequenced are collected. For the taxonomic assignment, the operator is queried about the database and taxonomic algorithm to be used.

Pre-processing and ASV formation: The second module was created using Python and R. Python is utilized to call the R script when the user desires to carry out pre-processing and ASV formation. The R script leverages the DADA2 package (version 1.30.0)[5] from the Bioconductor repository for this specific stage. The script generates a quality report for the raw reads. Subsequently, it performs read trimming and filtering, followed by the generation of a new quality profile for the filtered reads. During this filtration and trimming phase, ten base pairs are removed from the 5' end of all reads. Only reads exceeding a minimum length of two hundred base pairs, devoid of undetermined nucleotides ("N"), and possessing maximum expected error rates of one and two for forward and reverse reads, respectively, are retained for downstream analysis. Sequencing error plots are then generated to visualize the distribution of actual errors compared to a theoretical model. This theoretical error model serves as the foundation for ASV creation. Two output files are generated: a FASTA file containing the reference sequences for each ASV and an Excel spreadsheet summarizing the read counts for each ASV across all samples analyzed. Additionally, the script provides a report detailing the number of reads remaining after each processing step.

Taxonomic Assignment: Building upon the preceding module, taxonomic assignment leverages both Python and R. Python facilitates the seamless integration of the R script within the analysis pipeline. This stage utilizes the DECIPHER package (version 2.30.0) [18] from the Bioconductor repository, providing a comprehensive suite of functions for sequence classification based on taxonomic lineage. Crucially, taxonomic assignment necessitates a well-defined training set that links 16S rRNA gene sequences to their corresponding taxonomies. For this purpose, the script employs two established reference databases: SILVA [19], a database known for its extensive size, rigorous curation, pre-aligned sequences, and focus on detailed taxonomic classification, and the Ribosomal Database Project (RDP) [20] a curated database, while smaller in size, offers less frequent updates but is well-suited for routine amplicon analysis workflows due to its emphasis on established taxa. The script generates an Excel file summarizing the taxonomic classification for each ASV, along with the associated confidence scores for each classification.

Results exploration: The final module leverages the R package phyloseq (version 1.46.0) [21] from the Bioconductor repository for results exploration and visualization. This package offers functionalities for identifying unclassified Amplicon Sequence Variants that warrant further investigation.

Phyloseq enables the generation of alpha diversity plots to assess the biodiversity within each analyzed sample. These plots typically utilize metrics such as Observed richness, Chao1 richness estimator, Shannon diversity index, and Fisher's alpha diversity index. Following the removal of singletons (reads occurring only once), the script normalizes read counts either through rarefaction or by calculating relative frequencies. In addition, the beta diversity is also generated using PCoA and NMDS methods in order to analyze biological diversity between samples

To facilitate visualization of taxonomic abundance within samples, bar plots are generated for each taxonomic class. This graphical representation allows researchers to explore and interpret the distribution of taxa across samples. The integration of this module with the preceding ones is achieved through the utilization of Python.

3 Results and discussion

A critical step in amplicon analysis accuracy is quality control and read filtering. Overly stringent filtering parameters can lead to excessive read removal, potentially compromising the sensitivity of the analysis, particularly for low-abundance taxa. Conversely, less stringent parameters, while potentially beneficial for datasets with higher sequencing error rates, may introduce additional errors into the analysis. These errors can be misinterpreted as biological variation downstream, leading to false results [11].

As illustrated in Table 1, the utilization of the same filtering parameters, during the DADA2 module, may yield different high-quality, ready to analyze, dataset fractions, as a consequence of the overall quality of the produced raw sequencing data. For instance, while for dataset ATCC2 only 43 percent (input/nochim) of the original raw data was retained (after chimera removal: nochim), whereas, in the case of Zymo2, about 77 percent was retained.

Table 1. Overview of DADA2 module output.

ID sample	input	filtered	denoisedF	denoisedR	merged	nonchim
ATCC1	81242	36478	36359	36344	36159	35806
ATCC2	37746	16385	16303	16291	16226	16213
ATCC3	141119	119989	119764	119807	119067	116068
Zymo1	182644	144513	144264	144193	137987	136072
Zymo2	217600	175028	174829	174735	167517	167238

Often, to tackle such issues in input data, a detailed inspection of dataset quality trends is advised. In Figure 3 (appendix) it is depicted, as an example,

the raw quality profiles (upper panel) for both forward and reverse reads of sample ATCC2. A progressive decline in quality scores is evident with increasing sequencing cycle numbers, indicating a decrease in base call accuracy. Additionally, the reverse reads exhibit a consistently lower quality profile compared to the forward reads. This observation aligns with documented limitations of the Illumina sequencing platform regarding read quality towards the 3' end [22]. The lower panel of Figure 3 presents the quality profiles following read filtering. A marked improvement in quality scores is observed for both forward and reverse reads. However, the reverse reads continue to display greater variability in quality, particularly at the 3' end. This suggests that a targeted trimming strategy focusing on the 3' end might be beneficial. However, it is crucial to acknowledge the potential drawbacks of excessive trimming as gene length can be a taxonomic characteristic since the result of evolutionary processes such as mutations and horizontal gene transfer can lead to variation in gene length [23]. Furthermore, since the data originates from paired-end sequencing, a sufficient overlap between forward and reverse reads is essential. This overlap (merge) plays a critical role in the construction of a high-fidelity amplicon library and facilitates the identification of chimeric sequences, which can arise from artifacts during library preparation [24]. Excessive trimming could potentially compromise this overlap, hindering downstream analyses.

Upon proper dataset filtering, denoising, and chimera removal, the DADA2 module proceeds with the clustering of amplicon sequence variants (ASV) approach. Using our test data we obtained 86 raw ASVs which were classified using either Silva or RDP databases, resulting in 75 or 67 classified ASVs, respectively, as depicted in Table 2. Since our mock datasets only included well-known taxa, we agglomerated all ASVs classified up to genus level, to evaluate the performance of both classifiers. Although the results were fairly consistent, they suggested that, with our test data, the RDP classification database performed about 10 percent better (18 vs 16 classified ASVs). Thus, the choice of reference database significantly impacts taxonomic assignment. When using the RDP database, nineteen out of eighty-six identified ASVs lacked matches. This number dropped to eleven using the more comprehensive SILVA database. This discrepancy can result from inefficiencies in read filtration and chimera or due to contaminations during the experimente since unidentified ASVs, were not even detected as bacteria.

Table 2. Number of ASVs classified by RDP or SILVA according to filtering parameters. Filtered (NA)- Without taxonomic classification; Genus ASVs-Agglomeration of ASVs classified to genus level

Database	**Raw**	**Filtered (NA)**	**Genus ASVs**
RDP	86	67	18
SILVA	86	75	16

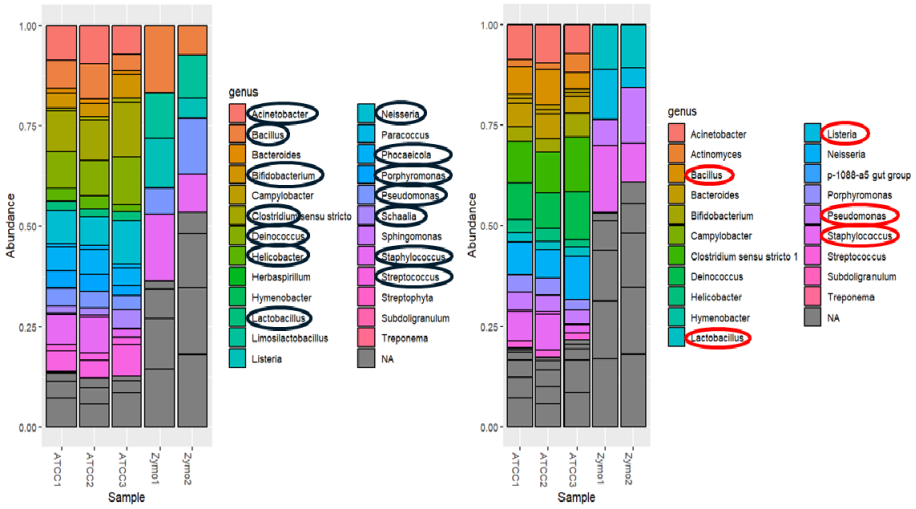

Fig. 1. Genus classification: the left side corresponds to the classification using the RDP database while the right side corresponds to the classification using the SILVA database. The blue ovals identify the organisms found in the ATCC community and the red the microorganisms found in the Zymo mock community.

Figure 1 presents an exploration of the taxonomic results at the genus level. A noteworthy observation is the high prevalence of unassigned reads (represented as "NA"), especially for sample Zymo 2, reaching up to 50 percent of the filtered reads. This finding warrants an investigation into the potential sources of this issue. Ideally, a successfully classified ASV should have a designated genus. Unassigned reads could stem from the implemented bioinformatic pipeline, particularly the pre-processing and filtration steps, which might be overly permissive, allowing reads with low quality and high error rates to pass through the filtering stages. This necessitates a reevaluation of the filtration criteria employed. Furthermore, inefficient chimera identification and removal algorithms could contribute to unassigned reads. Finally, issues during the experimental phase, such as DNA extraction, PCR amplification, or sequencing errors, could also lead to unassigned reads. Determining the most likely cause requires further investigation, although a combination of experimental errors and bioinformatic bias is a strong possibility.

Analyzing, more in detail, the RDP database results reveals that only four out of twenty expected organisms were not identified from the ATCC community, while only half were detected in the Zymo community. Notably, three of the unidentified organisms were abundant in the sample, while the others were low abundance. This suggests that the filtration parameters might be too stringent, hindering the detection of low-abundance taxa. Furthermore, the analysis identified eleven genera not present in either reference community. These could potentially be contaminants introduced during the experimental procedures. Similar

results were obtained with the SILVA database, with the exception that Bacillus was not identified in the Zymo samples, further increasing the number of unassigned reads. This highlights the crucial role of the reference database in influencing results since databases can introduce bias and potentially lead to classification errors due to the propagation of taxonomic inaccuracies present within the database itself.

Nonetheless, depending on the targeted biological questions, the exploitation module, based on the Phyloseq package, allows to infer several critical aspects of microbial communities. Figure 2 exemplifies two key components of this exploitation package: Inference of alfa diversity indices (providing crucial information, for instance, sample replicate evenness and diversity) and Inference of sample-to-sample clustering features (beta diversity) according to existing metadata.

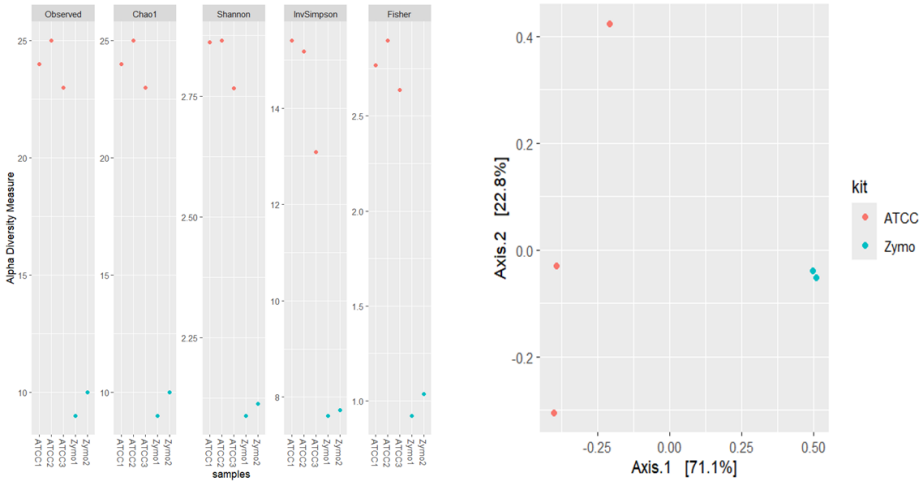

Fig. 2. Examples of output from phyloseq data exploitation. Left-Alpha diversity; Right-Beta diversity for each sample. The color scheme used to distinguish mock community commercial kit

More in detail, in Figure 2, the alfa diversity (left side) shows that ATCC communities are constituted by a higher number of microorganisms when compared with Zymo as expected from the theoretical composition of each mock. In addition, the beta diversity (right side) through the method PCoA shows that the two principal components are enough to explain about 94 percent of the variance observed within the samples, showing also that both mocks can be grouped in two independent groups that correspond to each mock. Furthermore, the samples from each mock are relatively close in the space displaying homogeneity within samples of the same mock

4 Conclusion and Future Perspectives

The application of the personalized scripts revealed a significant proportion of unassigned reads, particularly in sample Zymo 2. This finding suggests potential issues with the bioinformatic pipeline, including overly permissive pre-processing and filtration steps, or inefficient chimera identification algorithms. Additionally, experimental errors during DNA extraction, PCR amplification, or sequencing could contribute to unassigned reads. Further investigation is necessary to pinpoint the exact cause, but a combination of experimental and bioinformatic factors is likely. This conclusion highlights the importance of quality control and read filtering in maintaining data integrity while acknowledging the trade-off between sensitivity and accuracy. Stringent filtering parameters can eliminate low-abundance taxa, whereas overly permissive filtering might introduce noise and errors.

The choice of the reference database significantly impacted taxonomic assignment. While the SILVA database provided more comprehensive results compared to the RDP database, a substantial number of unassigned reads remained. These unassigned reads could potentially represent contaminants introduced during the experimental procedures.

The results also highlighted limitations associated with the 16S rRNA gene for species-level classification. While the V4 region offers sensitivity up to the genus level, employing complementary regions, such as the V3 region, might be necessary for more precise species identification. However, even the V3 region has limitations, as evidenced by recent studies emphasizing the challenges of using the 16S rRNA gene as a reliable species-specific marker.

The developed scripts demonstrate the potential for accurate amplicon analysis using the test datasets. However, to ensure generalizability and applicability to real-world studies, further validation and benchmarking are necessary. The results also highlight areas for improvement to achieve more precise analyses and minimize false positives, which could lead to erroneous interpretations in practical applications.

Extending support for additional sequencing technologies and incorporating established tools like USEARCH/VSEARCH [6], QIIME2 [25], RDP classifier [26], and SINTAX [27] would enhance the robustness and comprehensiveness of the analysis pipeline. Furthermore, the ability to customize amplicon analysis parameters would be valuable, as each dataset possesses unique characteristics and might require adjustments to processing parameters.

The integration of machine learning algorithms and filtration parameter prediction tools, such as FIGARO [28], holds promise for automating read filtration optimization, significantly reducing the time required to optimize and execute a complete amplicon analysis while promoting accurate results.

References

1. Committee on Science Needs for Microbial Forensics: Developing an Initial International Roadmap. Board on Life Sciences. Division on Earth and Life Studies. National

Research Council. Science Needs for Microbial Forensics: Developing Initial International Research Priorities. Washington (DC): National Academies Press (US). 2014. Microbial Science: Ecology, Diversity, and Characterizing the Microbial World.
2. Fakruddin M, Mannan KSB, Mazumdar RM, Chowdhury A, Hossain MN. Identification and characterization of microorganisms: DNA-fingerprinting methods. Songklanakarin J. Sci. Technol. 2013.
3. Yadav BS, Ronda V, Vashista DP, Sharma B. Sequencing and computational approaches to identification and characterization of microbial organisms. Biomed Eng Comput Biol. 2013. https://doi.org/10.4137/BECB.S10886.
4. Hakimzadeh A, Abdala Asbun A, Albanese D, Bernard M, Buchner D, Callahan B, Caporaso JG, Curd E, Djemiel C, Brandström Durling M, Elbrecht V, Gold Z, Gweon HS, Hajibabaei M, Hildebrand F, Mikryukov V, Normandeau E, Özkurt E, M Palmer J, Pascal G, Porter TM, Straub D, Vasar M, Větrovský T, Zafeiropoulos H, Anslan S. A pile of pipelines: An overview of the bioinformatics software for metabarcoding data analyses. Mol Ecol Resour. 2023. https://doi.org/10.1111/1755-0998.13847.
5. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016. doi: 10.1038/nmeth.3869.
6. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016. https://doi.org/10.7717/peerj.2584.
7. OLIVEIRA, R. R. M. et al. PIMBA: A PIpeline for MetaBarcoding Analysis. Advances in Bioinformatics and Computational Biology. 1ed. Switzerland: Springer, 2021, v. 13063, p. 106–116, 2021.
8. Callahan BJ, Sankaran K, Fukuyama JA, McMurdie PJ, Holmes SP. Bioconductor Workflow for Microbiome Data Analysis: from raw reads to community analyses. F1000Res. 2016. https://doi.org/10.12688/f1000research.8986.2.
9. Ha SM, Kim CK, Roh J, Byun JH, Yang SJ, Choi SB, Chun J, Yong D. Application of the Whole Genome-Based Bacterial Identification System, TrueBac ID, Using Clinical Isolates That Were Not Identified With Three Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) Systems. Ann Lab Med. 2019. https://doi.org/10.3343/alm.2019.39.6.530.
10. Coissac E, Riaz T, Puillandre N. Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol. 2012. https://doi.org/10.1111/j.1365-294X.2012.05550.x.
11. Alexander M Piper. Jana Batovska. Noel O I Cogan. John Weiss. John Paul Cunningham. Brendan C Rodoni. Mark J Blacket. Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience. Volume 8. Issue 8. 2019. https://doi.org/10.1093/gigascience/giz092.
12. Colovas J, Bintarti AF, Mechan Llontop ME, Grady KL, Shade A. Do-it-Yourself Mock Community Standard for Multi-Step Assessment of Microbiome Protocols. Curr Protoc. 2022. https://doi.org/10.1002/cpz1.533.
13. Higashi S, Barreto Ada M, Cantão ME, de Vasconcelos AT. Analysis of composition-based metagenomic classification. BMC Genomics. 2012. https://doi.org/10.1186/1471-2164-13-S5-S1.
14. Zhang T, Li H, Ma S, Cao J, Liao H, Huang Q, Chen W. The newest Oxford Nanopore R10.4.1 full-length 16S rRNA sequencing enables the accurate resolution of species-level microbial community profiling. Appl Environ Microbiol. 2023. https://doi.org/10.1128/aem.00605-23.
15. Compson ZG, McClenaghan B, Singer GA, Fahner NA, Hajibabaei M. Metabarcoding From Microbes to Mammals: Comprehensive Bioassessment on a Global Scale. Frontiers in Ecology and Evolution. 2020. https://doi.org/10.3389/fevo.2020.581835.

16. López-Escardó D, Paps J, de Vargas C, Massana R, Ruiz-Trillo I, Del Campo J. Metabarcoding analysis on European coastal samples reveals new molecular metazoan diversity. Sci Rep. 2018. https://doi.org/10.1038/s41598-018-27509-8.
17. Bik HM. Just keep it simple? Benchmarking the accuracy of taxonomy assignment software in metabarcoding studies. Mol Ecol Resour. 2021. https://doi.org/10.1111/1755-0998.13473.
18. ES Wright (2016) "Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R." The R Journal, 8(1), 352-359.
19. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013. https://doi.org/10.1093/nar/gks1219.
20. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014. https://doi.org/10.1093/nar/gkt1244.
21. McMurdie PJ, Holmes S. Phyloseq: a bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. Pac Symp Biocomput. 2012.
22. Kwon S, Park S, Lee B, Yoon S. In-depth analysis of interrelation between quality scores and real errors in Illumina reads. Annu Int Conf IEEE Eng Med Biol Soc. 2013. https://doi.org/10.1109/EMBC.2013.6609580.
23. Bartoš O, Chmel M, Swierczková I. The overlooked evolutionary dynamics of 16S rRNA revises its role as the "gold standard" for bacterial species identification. Sci Rep. 2024. https://doi.org/10.1038/s41598-024-59667-3.
24. Werner JJ, Zhou D, Caporaso JG, Knight R, Angenent LT. Comparison of Illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. ISME J. 2012. https://doi.org/10.1038/ismej.2011.186.
25. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, and Caporaso JG. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology. 2019. https://doi.org/10.1038/s41587-019-0209-9
26. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007. https://doi.org/10.1128/AEM.00062-07.
27. Edgar RC. SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. bioRxiv; 2016. https://doi.org/10.1101/074161.

28. Michael M. Weinstein, Aishani Prem, Mingda Jin, Shuiquan Tang, Jeffrey M. Bhasin bioRxiv. 2019. https://doi.org/10.1101/610394

Appendix

Table 3. ATCC mock community constitution and relative abundance.

Organism	**Abundance (percentage)**
Acinetobacter baumannii	5
Bacillus pacificus	5
Phocaeicola vulgatus	5
Bifidobacterium adolescentis	5
Clostridium beijerinckii	5
Cutibacterium acnes	5
Deinococcus radiodurans	5
Enterococcus faecalis	5
Escherichia coli	5
Helicobacter pylori	5
Lactobacillus gasseri	5
Neisseria meningitidis	5
Porphyromonas gingivalis	5
Pseudomonas paraeruginosa	5
Cereibacter sphaeroides	5
Schaalia odontolytica	5
Staphylococcus aureus	5
Staphylococcus epidermidis	5
Streptococcus agalactiae	5
Streptococcus mutans	5

Table 4. Zymo mock community constitution and relative abundance.

Organism	**Abundance (percentage)**)
Listeria monocytogenes	12
Pseudomonas aeruginosa	12
Bacillus subtilis	12
Escherichia coli	12
Salmonella enterica	12
Lactobacillus fermentum	12
Enterococcus faecalis	12
Staphylococcus aureus	12
Saccharomyces cerevisiae	2
Cryptococcus neoformans	2

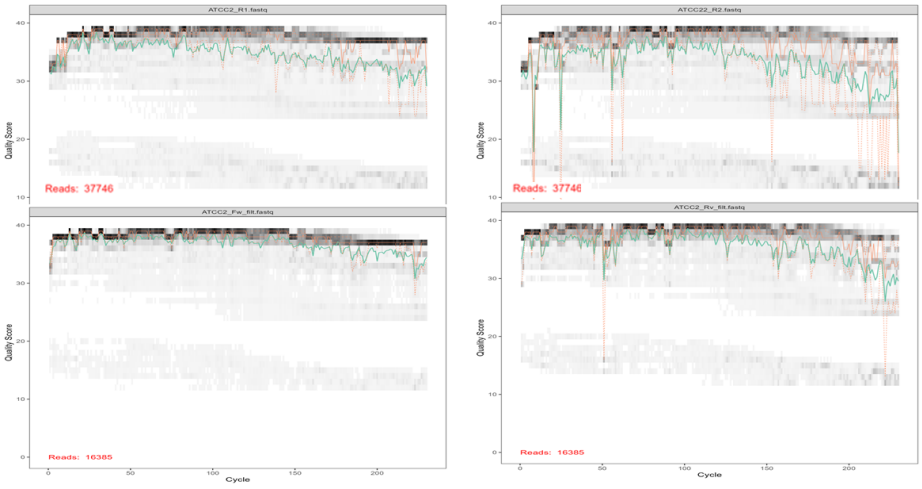

Fig. 3. Quality profile for the sample ATCC2. Top- unfiltered; Bottom- filtered.