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. Our project aims to assess metabarcoding tools across diverse sequencing platforms using mock community datasets. We'll evaluate each workflow phase and develop adaptable scripts integrating optimal tools for each phase. We'll apply these scripts to a real-world case study within our laboratory setting. Our workflow addresses the complexity of selecting suitable algorithms, particularly in mitigating errors and biases. We'll examine diverse methods for taxonomic assignment and ensure robustness through benchmark studies.

Keywords: Metabarcoding

1 Objective

Our project aims to evaluate the performance of metabarcoding analysis tools using datasets generated with different sequencing platforms (Illumina, PacBio, and Oxford Nanopore) systematically. We intend to assess each step of the analysis workflow using various tools such as DADA2 or USEARCH/VSEARCH, and different versions of each tool to determine the most accurate approach for every phase of the workflow. For this purpose, we will resort to commercial mock community sequencing datasets (e.g. Zymobiomics and ATCC) and customized mock communities created in our laboratory, which facilitates bioinformatic performance assessment

We intend to develop 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 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, for species identification. However, these evolutions increased massively the biological data produced [4] and, according to the National Human Genome Research Institute, is expected to generate between 2 to 40 exabytes of data within the next decade.

In response, a vast diversity of bioinformatics tools have been designed to allow the analysis of metabarcoding data. QIIME 1 and USEARCH are among the first highly successful software that can be combined to create metabarcoding data analysis pipelines initially designed for bacterial 16S rRNA amplicon analysis. However, the applications of metabarcoding have been expanded to a wide diversity of taxa from the most diverse environments, resulting in a drastic 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, and USEARCH/VSEARCH while, the second is a set of pipelines providing 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, meanwhile crucial for data analysis and results interpretation, 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 the sequencing of amplicon data and are not suitable for data generated from other approaches, such as whole-shotgun [5]. Nonetheless, in recent years, a new category of algorithms to analyze genomic similarity is gaining relevance allowing analysis of samples beyond taxonomic profiling, such as metabolic pathway analysis [6].

A standard workflow for amplicon data is followed to perform metabarcoding analysis varying the algorithm choice depending on the characteristics of the dataset [4]. In this article, each phase will be analyzed along with the problems and biases faced and how the different algorithms overcome them. It is impor-

tant to note that the workflow can suffer changes according to the experimental procedure implemented and the aim of the study.

3 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.

To process the amplicon sequencing data from Nanopore, the first step is to assign each sequence back to its respective source sample based on the unique oligonucleotide present in the sequencing adapters. To achieve this, software such as bcl2fastq, cutadapt or lima demultiplexes the sequences into individual FASTQ files that correspond to their respective samples allowing posterior analysis [4, 7].

Sequencing adapters and other non-biological information such as primers must be removed with tools like Trimmomatic or AdapterRemoval to eliminate nucleotides added to the target gene during the amplification and sequence process as they do not provide useful information and can lead to wrong conclusions [4].

During the practical procedure, 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 normally biased resulting in erroneous conclusions [8]. To remove sequencing errors software, such as DADA2, has a sequence filtration algorithm usually based on the read quality score. Nevertheless, this is coarse filtering where parameters such as read length, number of ambiguous bases, or average quality score should be carefully considered, especially for higher error sequencing platforms like Oxford Nanopore. Excessively conservative parameters can result in the removal of too many reads and therefore loss of sensitivity to low-abundance taxa introducing bias to the output [7].

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 creating false positive results [4]. When the amplicon data is paired-end, the overlapping sequences usually may be merged before or after the quality filtration step with some exceptions where the merging process must be done after the clustering process [4].

After the sequence filtration and artificial sequence removal, 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 that originate noise and can

be misinterpreted as biological variation substantially increasing downstream computation [7, 9]. Nonetheless, these denoising algorithms tend to discard low-abundance sequences which are likely to be artifacts, however, in some scenarios, is hard to distinguish noise from a real low-abundance signal and therefore discarding rare taxa. This problem gains special importance when the dataset has relatively low sequencing depth [4].

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, creating false negative results [4, 7].

Standard workflows depended on clustering using OTUs, however, this approach does not take into consideration individual sequence quality information and statistical information, and therefore, using an arbitrary threshold can add bias to the results because error rates often vary substantially between sequencing runs and PCR protocols [5]. To solve this problem, new denoising algorithms that use statistical models and incorporate sequence quality were developed to distinguish true biological sequences from sequencing noise and correct single nucleotide differences. The clustering of amplicon sequence variants (ASV) is a new denoising algorithm that is present in DADA2 and 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 [9]. A study done in 2021 reported that ASV provides a biologically informative fine-scale resolution compared to OTU clustering [4, 7].

On the other hand, 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, however, removal of non-specific amplification products is challenging and usually requires manual curation, although when using rRNA as the target gene, secondary structure prediction can be used to ensure that sequences do not contain considerable variation in highly conserved genes [7]. De novo method can also discard sequences incorrectly flagged as chimeric creating bias, especially for data sets with low sequence depth. In an attempt to recover false positive chimeras, some algorithms, such as NextITS, verify if the sequence is present in multiple samples since the formation of identical chimeras across samples is highly unlikely [4]. It is also important to note that the error removal tools sometimes are specific for a certain sequence method and cannot be used extensively [5].

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 isn't a unanimous decision. However, the use of negative controls can be useful to eliminate contaminations 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 technical bias linked to experimental procedures [4, 10].

4 Taxonomic assignment

After assuring that the data meets the quality standards and pre-processing is done, 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 method 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 and, in the worst scenario, over-classification can lead to false-positive results by classifying an unsequenced organism [7].

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 [11]. 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 [12] are the most widely used sequence databases for DNA metabarcoding analysis. However, the coverage is incomplete and uneven with extreme variation based on the taxonomic group [13]. For instance, the study done by López-Escardó et al (2018) [14] 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. To realize these studies mock communities containing a known heterogeneous composition including varying numbers of taxonomic groups and individuals per taxa should be used to mimic

the natural ecosystem from where the samples were obtained [10]. 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 don't meet the quality standards defined, they can be used to test modifications, such as different pre-processing algorithms and sequence filtering approaches and parameters, to optimize the pipelines for an accurate and precise metabarcoding analysis [15]. Although mock communities are necessary to evaluate the bioinformatics performance and the experimental protocols, they also can impose bias if not well-chosen since mock communities that better resemble with the environment in study are necessary to more precisely access every phase. Since most of the commercially available mocks are mainly focused on the human microbiome they may not have a relevant taxonomic representation for the environment of interest. For this scenario, the best approach is to use customized mocks aimed at the environment under study trying to reproduce it to have a relevant community to realize benchmark [10].

5 Project workflow

We will use the D6300 and D6305 mock communities from Zymo Research since these are established standards used extensively in metabarcoding analysis resulting in a wide availability of data produced by Illumina, PacBio, and Oxford Nanopore sequencing. In addition, we will also extend the analysis to other mock communities, including from ATCC and customized mock communities generated in our laboratory.

For the performance analysis of each dataset, we will use DADA2 [9] and USEARCH/VSEARCH [16] since these are the most common tools used to perform metabarcoding analysis. For each data set, two different versions of each software will be used to account for how the changes of each version will influence the final results. The organism will be classified using algorithms such as RDP classifier and SINTAX and the results will be evaluated by comparing with the information available about the mock communities, where the ability of the approach used to identify the organisms present will be assessed, along with the ability to quantify the proportion at which each organism occurs.

With the results retrieved from the performance study, a set of scripts will be developed where the information about the sequencing technique will be used to select the optimized analysis. However, the scripts will be designed for the user to have the freedom to choose which tool to use and will also be able to change the analysis parameters. Finally, the scripts developed will be applied and tested in a real case study, currently ongoing in the lab, where the microbes composition of marine sediments will be studied.

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, 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.

13. 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.
14. 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.
15. 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.
16. 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.