Hakimzadeh Ali (Orcid ID: 0000-0003-1336-7445)
Bernard Maria (Orcid ID: 0000-0001-9005-5563)
buchner dominik (Orcid ID: 0000-0002-8499-5863)
DJEMIEL Christophe (Orcid ID: 0000-0002-5659-7876)
Brandström Durling Mikael (Orcid ID: 0000-0001-6485-197X)
Elbrecht Vasco (Orcid ID: 0000-0003-4672-7099)
Hajibabaei Mehrdad (Orcid ID: 0000-0002-8859-7977)
Pascal Géraldine (Orcid ID: 0000-0002-5250-6594)
Vasar Martti (Orcid ID: 0000-0002-4674-932X)

A pile of pipelines: an overview of the bioinformatics software for metabarcoding data analyses

Ali Hakimzadeh[1], Alejandro Abdala Asbun[2], Davide Albanese[3], Maria Bernard[4,5], Dominik Buchner[6], Benjamin Callahan[7], J. Gregory Caporaso[8], Emily Curd[9], Christophe Djemiel[10], Mikael B. Durling[11], Vasco Elbrecht[12], Zachary Gold[13], Hyun S. Gweon[15, 16], Mehrdad Hajibabaei[22], Falk Hildebrand[17,18], Vladimir Mikryukov[1], Eric Normandeau[19], Ezgi Özkurt[17,18], Jonathan M. Palmer[20], Géraldine Pascal[21], Teresita M. Porter[22], Daniel Straub[23], Martti Vasar[1], Tomáš Větrovský[24], Haris Zafeiropoulos[25], Sten Anslan[1*]

[1] Institute of Ecology and Earth Sciences, University of Tartu, Estonia.
[2] Department of Marine Microbiology and Biogeochemistry, NIOZ Royal Netherlands Institute for Sea Research, Texel, Netherlands.
[3] Unit of Computational Biology, Research and Innovation Centre, Fondazione Edmund Mach, Italy.
[4] Université Paris-Saclay, INRAE, AgroParisTech, GABI, 78350, Jouy-en-Josas, France.
[5] INRAE, SIGENAE, 78350, Jouy-en-Josas, France.
[6] Aquatic Ecosystem Research, University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, Germany
[7] Department of Population Health and Pathobiology, College of Veterinary Medicine and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA.
[8] Center for Applied Microbiome Science, Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA.
[9] Vermont Biomedical Research Network, University of Vermont, Burlington, VT, USA.
[10] Agroécologie, INRAE, Institut Agro, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France.
[11] Department of Forest Mycology and Plant Pathology, Swedish University of Agricultural Sciences, Box 7026, 75007 Uppsala, Sweden.
[12] Aquatic Ecosystem Research, University of Duisburg-Essen, Universitaetsstrasse 5, 45141, Essen, Germany
[13] Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA[8]
[14] Southern California Coastal Watershed Research Project, Costa Mesa, CA, USA
[15] UK Centre for Ecology & Hydrology, Wallingford, Oxfordshire, OX10 8BB, UK.
[16] School of Biological Sciences, University of Reading, Reading, RG6 6EX, UK.
[17] Gut Microbes & Health, Quadram Institute Bioscience, Norwich Research Park, Norwich, Norfolk, NR4 7UQ, UK.
[18] Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UZ, UK.
[19] Institut de Biologie Intégrative et des Systèmes, Université Laval, Québec, QC, Canada.
[20] Center for Forest Mycology Research, Northern Research Station, US Forest Service, Madison, WI USA (current address: Genencor Technology Center, IFF, Palo Alto, CA USA)
[21] GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326, Castanet Tolosan, France.
[22] Department of Integrative Biology and Centre for Biodiversity Genomics, University of Guelph, Canada.
[23] Quantitative Biology Center (QBiC), University of Tübingen, Tübingen D-72076, Germany.
[24] Laboratory of Environmental Microbiology, Institute of Microbiology of the Czech Academy of Sciences, Vídeňská 1083, 14220 Praha 4, Czech Republic.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1755-0998.13847

[25] KU Leuven, Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Laboratory of Molecular Bacteriology, 3000 Leuven, Belgium.

§ Currently at NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA

* Corresponding author, sten.anslan@ut.ee

Keywords: metabarcoding, environmental DNA, bioinformatics, pipeline, amplicon data analysis, review

Abstract

Environmental DNA (eDNA) metabarcoding has gained growing attention as a strategy for monitoring biodiversity in ecology. However, taxa identifications produced through metabarcoding require sophisticated processing of high-throughput sequencing data from taxonomically informative DNA barcodes. Various sets of universal and taxon-specific primers have been developed, extending the usability of metabarcoding across archaea, bacteria, and eukaryotes. Accordingly, a multitude of metabarcoding data analysis tools and pipelines have also been developed. Often, several developed workflows are designed to process the same amplicon sequencing data, making it somewhat puzzling to choose one amongst the plethora of existing pipelines. However, each pipeline has its own specific philosophy, strengths, and limitations, which should be considered depending on the aims of any specific study, as well as the bioinformatics expertise of the user. In this review, we outline the input data requirements, supported operating systems, and particular attributes of thirty-two amplicon processing pipelines with the goal of helping users to select a pipeline for their metabarcoding projects.

Introduction

Advances in high-throughput sequencing (HTS) technologies have boosted the application of molecular methods for species identifications. Metabarcoding, the simultaneous tagging, sequencing, and identification of multiple species within a single environmental sample (Taberlet et al., 2012) is now a widely applied technique in biodiversity research (Compson et al., 2020). Metabarcoding involves PCR-based amplification of taxonomically informative gene fragments ('DNA barcodes', markers) that are subsequently sequenced to be used for species identifications in the presence of reference sequence data (DNA barcodes). Before identification, the sequencing data is processed in several steps (Fig. 1) where one of the first steps is usually performing quality control on the data. A sequence analysis pipeline is generated by applying various steps using a collection of software and algorithms with the ultimate goal of producing an accurate features table with potential taxon annotations by sample (i.e., with features metadata). In metabarcoding, features refer to amplicon sequence variants (ASVs), operational taxonomic units (OTUs), or annotated taxa; and their sample-wise distribution matrix can be further utilized in relevant biostatistical analyses.

With the emergence of practical guidelines (e.g., Bruce et al. 2021; Lear et al. 2018; Tedersoo et al. 2022), the scalability and throughput of environmental DNA (eDNA; a mixture of DNA from different organisms in an environmental sample; Taberlet et al., 2012) sample processing has contributed to the popularity of the metabarcoding approach among ecologists. However, one of the bottlenecks of metabarcoding is choosing how to process sequencing datasets into relevant feature tables bioinformatically. Among the first highly successful software developed for that purpose have been mothur (Schloss et al. 2009), USEARCH (Edgar, 2010), and QIIME 1 (Caporaso et al., 2010), which consists of algorithms that can be combined to create full metabarcoding data analysis pipelines. Over the years, these programs have been supplemented with additional algorithms to help reduce artifactual sequences and implement different sequencing clustering approaches. These pipelines were initially developed for microbial 16S rRNA amplicon analysis, but the applications of metabarcoding have been expanded to a wide range of taxa from various environmental samples, resulting in a boom in pipeline development. Some workflows include a set of newly designed algorithms, but others represent a combination of different open-source tools used for the different analysis steps bound into executable pipelines. From the lack of easy-to-use bioinformatics tools from the early age of metabarcoding, we have reached a phase where the choices are so numerous that it may be difficult to select among the multitude of analytical workflows.

Below, we delve into the properties of thirty-two software packages that can be used for the bioinformatics processing of metabarcoding data. In this review, we outline several key aspects of those metabarcoding software, including which ones represent software suites or pre-compiled pipelines, consideration of the software depending on the utilized sequencing platform, available operating system, and interface preference (Fig. 2). By addressing these components, we seek to offer a comprehensive understanding of the software landscape for metabarcoding projects. Since different users will have different needs, we do not seek to recommend the best-performing pipeline, a task that would be highly context-dependent, but rather to give an overview of the software that are available for metabarcoding data analysis.

Table 1 lists the software discussed here and their extended description and specific capabilities are outlined in Supplementary file 1.

Software suites and pre-compiled pipelines

The metabarcoding data processing software may be roughly divided into two categories based on their structure for executable algorithms – software providing a set of algorithms, and pipelines providing a pre-defined chain of algorithms. USEARCH, VSEARCH (Rognes et al., 2016), DADA2 (Callahan et al., 2016), OBITools (Boyer et al., 2016), mothur and QIIME 2 (Bolyen et al., 2019) are software suites that host numerous algorithms for sequence data analysis, thus are highly customizable to construct user-defined pipelines with a specific chain of commands and settings. VSEARCH largely mirrors the diverse functionalities of USEARCH, but without the requirement to purchase a license for a version that can handle large datasets and use more than 4 GB of a computer's memory. Besides consisting of a large set of unique data processing algorithms, mothur and QIIME 2 wrap some functionalities of VSEARCH and/or DADA2.

Software providing a pre-defined chain of algorithms represents full analytical pipelines with specific workflow steps, as depicted in Figure 1. The pre-defined pipelines consist of workflow steps validated on certain sequencing data to facilitate the metabarcoding data analysis, which may be especially convenient for users with few bioinformatics skills. Some workflows include a set of newly designed algorithms, but others represent a combination of different open-source tools used for the different steps that are bound into easily executable pipelines Although, these pipelines are pre-defined, they often allow the user to customize the settings depending on the characteristics of the sequencing data set.

Basic structure of a metabarcoding pipeline

Demultiplexing

Demultiplexed sequences are often provided to users since this process has been integrated into sequencing provider software, such as bcl2fastq (for Illumina raw data) and SMRT Tools (for Pacific Biosciences [PacBio] reads). Demultiplexing distributes the sequences into individual files, most often corresponding to the experiment's samples. However, when requesting multiplexed data (pooled sequences from multiple samples), the reads from a sequencing run may need to be demultiplexed before using some of the application software. The demultiplexing step is not incorporated into all software (Table 1). In cases where it is not, other programs such as cutadapt (Martin 2011), sdm (Falk Hildebrand et al. 2014), or lima (https://lima.how/; for single-end reads only) may be used. In cases where multiple markers are used per sample (via multiplex PCR), amplicons from different primer sets should also be split. The latter step is included in the Anacapa and VTAM pipeline (Curd et al., 2019; González et al., 2020), where different markers are automatically separated based on the primer sequences.

Primer trimming

Sequencing adapters, indexes, and primers should be removed before the following analyses. Depending on the data structure, the former two may be absent, but the programs mentioned above (for demultiplexing) may be used to double-check this and remove primers. Adapter/primer clipping is often implemented into a pipeline by wrapping the cutadapt, Trimmomatic (Bolger et al., 2014) or AdapterRemoval (Lindgreen, 2012) functionality, in others done during quality filtering and demultiplexing steps (e.g., sdm; Table S1).

Quality filtering & merging paired-end reads

The following phases of a standard DNA metabarcoding pipeline is sequence filtration based on the read quality scores, removal of putative chimeric/artifactual sequences, the definition of features (e.g., ASVs, OTUs), and taxonomic annotation of the features (Fig. 1). In the case of paired-end data, the merging process of the overlapping sequences may be performed before or after the quality filtering step and even sometimes after the sequence clustering step. There are a multitude of strategies for performing the above-listed processes, where the selection of an approach may depend on the specific characteristics of the sequencing data or the aims of the study. The strategies for quality filtering include per-sequence or per-nucleotide(s) based filtering. Per-sequence filtering includes discarding the whole sequence if it does not meet the threshold requirements, whereas the per-nucleotide(s) approach truncates the sequence from the position below the threshold to keep a partial amplicon. Among the quality threshold calculation methods, the filtering based on the expected number of errors (sum of the error probabilities) is preferred over the average quality score threshold (Edgar & Flyvbjerg, 2015), because a 'good' average quality score may mask several bases with relatively high error probabilities that can subsequently propagate into false positive features. Haplotype-level (ASV) analyses may require relatively stringent quality cutoffs for accurate fine-resolution analyses, whereas cutoffs may be more lenient when generating OTUs (because clustering collapses many of the accumulated errors during sequencing) or if summarizing data to more inclusive taxonomic ranks (species, genera, etc.).

Artifacts filtering

Putative chimeric sequences are most commonly removed by comparing sequences against each other (*de novo* method), but with the existence of an appropriate (curated, chimera-free) reference database, additional reference-based chimera filtering is recommended (Tedersoo et al. 2022). *De novo* methods tend also to discard sequences that are incorrectly flagged as chimeric (false-positive chimeric sequences; Pauvert et al., 2019; Tedersoo et al., 2022). The loss of these false positive chimeric sequences detection may be more 'costly' for datasets with low sequencing depths. To attempt to rescue those real members of the sequenced community (false-positive chimeras), NextITS (Mikryukov et al., 2022) and FROGS (Bernard et al., 2021; Escudié et al., 2018) pipelines have implemented an approach to recover sequences that occur in multiple samples (because the formation of an identical chimera in different PCR runs is highly unlikely). With NextITS, it is also possible to inspect the distribution of UCHIME scores (Edgar et al., 2011) of putative chimeras, which allows adjustment of sensitivity-specificity tradeoffs in chimera discrimination according to the study aims. A custom false-positive chimeras recovery method is also implemented in BIOCOM-PIPE

(Djemiel et al., 2020), where initially discarded chimeras can be recovered based on their taxonomic assignments.

Denoising & clustering

The formation of features in many pipelines includes both ASVs and OTUs (Table 1). ASVs are identical denoised reads with as few as 1 base pair difference between variants, representing an inference of the biological sequences prior to amplification and sequencing errors (Callahan et al., 2017). ASVs are mainly formed through the two most popular denoising algorithms, DADA2 and UNOISE (Edgar, 2016). Although features formed via UNOISE are referred as zOTUs (zero-radius OTUs; Edgar, 2016), sometimes also as ESVs (exact sequence variants; Buchner et al., 2022; Porter & Hajibabaei, 2022), we herein denote those with a unified term – ASVs. Although less frequently implemented in the pipeline, deblur (Amir et al. 2017) and obiclean (Boyer et al., 2016) are other denoising algorithms for ASVs formation (Table S1). The OTU clustering approaches include a much wider set of algorithms across different software (Table S1), which typically rely on global sequence similarities. Notably, the clustering process in SCATA (Durling et al., 2011) includes collapsing of homopolymer regions to account for homopolymer-length errors during sequencing (which are especially common on 454 and Ion Torrent platforms; Laehnemann et al., 2015). Similarly, before the formation of features, NextITS implemented the correction of homopolymer errors in PacBio reads. Swarm (Mahe et al., 2022) is a notably different sequence clustering approach. It relies on the maximum number of differences between reads (local linking threshold), where clusters are resilient to input-order changes, therefore forming stable, high-resolution features (herein referred to as swarm-clusters).Swarm is currently implemented in Cascabel, CoMA, FROGS, LotuS2 (Özkurt et al., 2022), MICCA (Albanese et al., 2015), NextITS and PEMA (Zafeiropoulos et al., 2020) pipelines (Table S1; Supplementary File 1).

Which type of features to prefer may be context-dependent, and both may even be used in the same study. Denoised ASVs provide a biologically informative fine-scale resolution that collapsed during the OTU formation process (Callahan et al., 2016). For example, by testing several ASVs and OTUs-based workflows for detecting the *Botrylloides* (Ascidiacea) haplotypes, Couton et al. (2021) reported that ASVs pipeline (DADA2) retrieved all expected haplotypes, whereas OTUs datasets (99.5% threshold for clustering) missed several expected haplotypes by collapsing very closely related ones into a single OTU. By default, denoisers tend to discard low-abundant sequence variants, which are more likely to be artifacts(Anslan et al., 2021; Reitmeier et al., 2021). Although denoising greatly lowers the fraction of spurious features (e.g., De Santiago et al., 2021), in some contexts it may be difficult to separate noise from a real signal in low abundant ASVs. For example, the denoising process might discard some rare taxa, i.e., ASVs with a low number of sequences (Edgar, 2016; Nearing et al., 2018). This may have a larger impact when working with a data set with a relatively low sequencing depth. Nevertheless, in some pipelines (e.g., DADA2, FROGS, VSEARCH, and USEARCH), the sensitivity to rare ASVs can be modified according to the user's needs. Importantly, ASVs represent stable and reproducible units across studies whereas OTUs are dataset-specific features (Callahan et al., 2017). However, the ASVs approach may not accurately reflect species composition in the community of e.g. metazoans with highly variable levels of intraspecific polymorphism in the COI gene (Brandt et al., 2021) and fungi with multiple

different ITS copies per genome and their size polymorphism (Tedersoo et al. 2022; Estensmo et al. 2021) except when the treatment of ITSs is particularly taken into account (as e.g. in FROGS; (Bernard et al. 2021).If relevant, upon formation of ASVs, those may be subjected to further clustering (Antich et al., 2021; Brant et al., 2021; Porter & Hajibabaei, 2020). The latter approach is implemented in e.g. MetaWorks (Porter & Hajibabaei, 2022), PipeCraft2 (Anslan et al., 2017), and dadasnake (Weißbecker et al., 2020). Additionally, QIIME 2, nf-core/ampliseq (Straub et al. 2020, Ewels et al 2020) and LotuS2 support the features collapsing by annotated taxon levels, resulting in taxa features. Overall, the resulting community patterns of a study are often highly similar regardless of the utilized feature (e.g., Glassman & Martiny, 2018; Kang et al., 2021; Porter & Hajibabaei, 2020), but may vary in recovering rare taxa (Nearing et al., 2018).

After the formation of features, the presence of a long tail of low-abundant units is common. This tail is often discarded, assuming that a large proportion of low-abundant features are artefactual (Reeder & Knight, 2009; Huse et al., 2010). However, without applying arbitrary cutoff levels (e.g., removing features with <10 reads per-sample; Brown et al., 2015), the post-clustering process aids in removing the erroneous features but keeping the rare, potentially real ones. Post-clustering tools, such as LULU (Froslev et al., 2017) are implemented in AMPTk (Palmer et al., 2018), eDNAflow (Mousavi-Derazmahalleh et al., 2021), APSCALE (Buchner et al., 2022), LotuS2, PipeCraft2, and ReClustOR (Terrat et al., 2019) in BIOCOM-PIPE.

Post-clustering, however, does not resolve the tag-switching phenomena, where some low-abundant non-artificial features may represent false-positive occurrences across samples. Tag-switching is a well-documented issue (e.g., Carlsen et al., 2012; Rodriguez-Martinez et al., 2022; Schnell et al., 2015), but is rarely considered in practice because the low proportions of tag-switching errors do not heavily impact the community-level analyses (e.g., Anslan et al., 2021). Nevertheless, the incorrect sample assignments of features artificially inflate the richness. For discarding potential tag-switching errors from the feature table, pipelines such as NextITS, LotuS2, and Dadaist2 (Ansorge et al., 2021) wrap the UNCROSS2 (Edgar, 2018) algorithm (from USEARCH). Based on the included control samples, AMPtk and VTAM attempt to automatically correct for tag-switching errors. Notably, the tag-switching issue can be minimized by accounting for this in the laboratory work protocol (Carøe & Bohmann, 2020; Taberlet et al., 2018). However, for further feature occurrence filtering to filter out low-confidence detections biological/technical replicates per sample are recommended (Gold et al., 2022). This allows examining the feature co-occurrence patterns across replicates to estimate detection probabilities and retain only high-confidence detections (by applying e.g., site occupancy modeling). Among the pre-compiled pipelines, VTAM implements a feature occurrence filtering procedure based on the user-defined number of technical replicates they appear in, and samples may be discarded when the sequence composition in the replicate samples is too dissimilar. Not incorporated to the pipelines discussed here, but the MetabaR package (Zinger et al., 2021) aids to detect different types of artifactual sequences, such as potential contaminants, tag-switches, and dysfunctional PCRs (on the basis of similarities between replicate samples).

Taxonomy assignment

In the reviewed pipelines, the most common taxonomy assignment methods include alignment-based (such as BLAST; Altschul et al. 1997) and sequence composition-based approaches (e.g., RDP Naïve Bayesian classifier; Wang et al 2007; see Table S1). Several studies have tested the accuracy of different taxonomy assignment methods (e.g., Edgar, 2018; Bokulich et al., 2018; Richardson et al., 2017; Curd et al., 2019; Hleap et al., 2021) and have recognized a relationship between the reference database completeness and the classification accuracy. Regardless of the taxonomic group, the reference databases are far from being complete (Gold et al., 2021; McGee et al., 2019; Nilsson et al., 2016; Weigand et al., 2019). Therefore, a trade-off between the detection of true-positives (correctly assigned sequences) and false-positives (incorrectly assigned sequences), i.e., the precision and the recall rate, should be considered when choosing a threshold for the classification (Edgar, 2018; Bokulich et al., 2018). Additionally, false negatives, which refer to unassigned sequences, should also be taken into account, as the trade-off between false positives and false negatives is particularly pertinent in this context. Hleap et al. (2021) suggested that a multilayer approach could enhance the effectiveness of similarity-based methodologies. The goal of this strategy is to improve the precision of taxonomic assignments, minimize the occurrence of false positives, and boost the efficiency of the classification process. VTAM's taxonomy assignment function has incorporated elements of this strategy. It begins the assignment process with a high percentage identity threshold, which is systematically lowered when there are not enough valid matches.

Although composition-based (and other 'complex') methods may be more sensitive to the patchy coverage databases than 'simple' alignment-based methods (Hleap et al., 2021), in certain circumstances Naïve Bayesian classifiers may outperform BLAST (Rosen et al., 2011). However, the assignment accuracy to higher taxonomic ranks (such as Family level) generally has similar performance across the approaches (Hleap et al., 2021). A recent development in QIIME 2 involves utilizing public microbiome data for probabilistic taxonomy assignment (Kaehler et al., 2019). This method offers several advantages, including the potential for higher resolution taxonomic classification for instance, it can enable species-level classification when previously only genus-level classification was possible. Other pipelines, such as LotuS2, can assign features from multiple taxonomic databases, to preferentially assign taxonomies based on databases that are specific to a given environment. FROGS returns an original multi-affiliation output to highlight databases conflicts and uncertainties taxonomic affiliations. AMPtk implements a hybrid taxonomy assignment that utilizes global alignment (VSEARCH) and SINTAX (Edgar, 2016) to calculate a consensus LCA (last common ancestor) taxonomy. Regardless of the taxonomy assignment methods used, the reference database should also include a proportion of non-target taxa (including potential contaminants) to limit the overclassification of features to the target taxa (e.g., Anslan et al., 2018).

Sequencing platform

The most commonly utilized high-throughput sequencing approaches for metabarcoding are short-read, second-generation technologies, such as those provided by Illumina platforms. These platforms produce a high number of high-quality paired-end short reads (up to 300 bp for single-end) with a relatively low cost per sample. Therefore, most

amplicon data analysis pipelines are set up to be able to handle paired-end sequencing data (Table 1, Fig. 2). As the MGI-Tech platforms may also produce paired-end reads (with comparable data quality and throughput properties compared with Illumina; Anslan et al., 2021), the paired-end compatible pipelines may be used to analyze data from the latter platforms as well.

Some pipelines are restricted to paired-end input, i.e. the analytical pipeline cannot be completed using only a single-end part of the data, or sequencing data from the long-read (third-generation) sequencing platforms (Table 1, Supplementary file 1). With the rapid developments in third-generation sequencing accuracy and throughput, there is increasing interest to generate longer metabarcodes, which potentially increases the taxonomic resolution (Tedersoo et al., 2021; Tedersoo et al., 2022) and has lower sequencing bias toward short amplicons (Castaño et al., 2020). Therefore, some software developed for short reads have been updated to also process longer sequences (specifically PacBio reads; Supplementary file 1). Although, some of the software considered here have performed well for sequencing data (HiFi reads) processing from PacBio platforms (e.g., Castaño et al., 2020; Heeger et al., 2018; Tedersoo & Anslan, 2019), the data from Oxford Nanopore Technologies (ONT) platform may require other customized approaches (Baloğlu et al., 2021). Herein listed software (Table 1) have not been specifically developed for analyzing ONT data, thus care should be taken when applying these tools for nanopore reads.

The sequencing depth may vary considerably between sequencing platforms. For example, Illumina MiSeq system may produce up to 25M 2x300 bp reads and NovaSeq up to 1600M 2x250 bp reads per flow cell, whereas the throughput of PacBio Sequel II(e) system is up to 4M HiFi reads. Since the denoising tools are sensitive (by default) to low abundant sequences, then one must be wary that strict denoising of low sequencing depth samples increases the number of false negatives, i.e., rare true positives may be denoised out (Furneaux et al., 2021), especially if the samples contain complex communities, such as found in soil. Besides sequencing depth, the detection of rare sequence variants may be affected by the different chemistry utilized by different platforms (e.g., NovaSeq vs. MiSeq; Singer et al., 2019). Importantly, denoising algorithms, such as UNOISE and deblur, are designed for Illumina reads and may not perform well with data from other sequencing platforms. Therefore, opting for an OTU clustering approach may be more appropriate for analyzing complex communities sequenced by the third-generation platforms. The remaining bioinformatically unscreened low-abundance spurious OTUs could then be abandoned after the post-clustering step, by e.g., filtering out unclassified features at the phylum level, and based on the number of samples they occur in (e.g., discard features only observed in one sample). Although, DADA2 has a specific denoising function to estimate errors from PacBio reads (Callahan et al., 2019) which performs well also on synthetic long reads (Callahan et al., 2021), its application may still require higher sequencing depth for high diversity samples (Furneaux et al., 2021). However, the throughput of the most recent PacBio long-read sequencing system, Revio (commercially available from the first half of 2023), is expectedly up to 15 times higher compared with Sequel II. But the performance of the denoisers with the greatly increased throughput of long-read data (exceeds the throughput of e.g., Illumina MiSeq) is yet to be tested.

In case the amplicon is shorter than the sequencing cycle (e.g., expected amplicon is ~130 bp, but one cycle synthesizes 250 bp), the Illumina NovaSeq and NextSeq platforms may extend the amplicon by adding a poly-G tail (with 'good' quality scores). Therefore, trimming primers from amplicon reads should be used by default, as this will discard the overhanging sequence parts. Fastp tool (Chen et al., 2018), wrapped also in PipeCraft2 and NextITS, may be used to specifically trim these non-biological poly-G (or poly-X) tails. Additionally, Phred scores from third-generation sequencing platforms range from 0-93, thus may require adjustments of the maximum quality score setting when using e.g., VSEARCH or USEARCH software (where the default is 41, for Illumina).

Operating systems and workflow managers

Unix-based operating systems (OS), such as Linux and macOS, are the most common and convenient platforms in bioinformatics, as users may run software with a comparable interface on a personal computer or high-performance computing (HPC) system. As a result, they are by far the most widely used for the development and use of bioinformatics tools. Accordingly, almost all the presented pipelines can be executed in Linux and/or macOS operating systems (Table 1; Fig. 2). Since many users have computers running on Windows-based operating systems, several pipeline developers have gone through the effort of making the Unix-based workflows executable in Windows; sometimes through native code adaptations or by making their software available in either containers or through websites (such as Galaxy; The Galaxy Community, 2022), making the pipelines independent of the OS (Table 1).

Some metabarcoding data analysis tools, such as DADA2 rely exclusively on R and are thus also compatible with any OS that can execute R. JAMP (https://github.com/VascoElbrecht/JAMP) is another R package that wraps full metabarcoding and haplotyping pipelines, although it is only available for Linux and macOS (Elbrecht et al., 2018). Additionally, with the development of containerization technology (e.g., Docker, Singularity), it becomes easier to develop bioinformatics pipelines that can run on the three major operating systems, Windows, macOS, and Linux. A container encapsulates the code and dependencies needed for the data analyses so that the pipeline may run reliably on any OS. Once the containerization software is installed, users are free to install all the underlying dependencies. For a few of the presented pipelines, the developers have included the pre-built containers and/or virtual machine images required to run it (Table 1, Supplementary file 1). Pipelines such as those distributed by nf-core/ampliseq, PipeCraft2, PEMA, and Tourmaline require utilizing Docker/Singularity containers at the back-end, so the core bioinformatics processes are running on a Linux environment but may also be executed on Windows and macOS systems. Moreover, containerized pipelines resolve the numerical instability issue occurring while running software on different computational platforms (Di Tommaso et al., 2017), ensuring the consistency of results and allowing more reproducible computational workflows.

Essentially all the pipelines can be run on any OS via containers or virtual machines. However, containers are preferred to virtual machines (e.g., VirtualBox), as virtualization (i.e., running a second OS on top of the main OS) has high overhead and comes at the cost of a computer's RAM usage, which ultimately limits the amount of data that can be processed.

Considering container engines, Docker is usually unavailable on HPC clusters, as potential vulnerability could provide means to gain root access to the system they are running on. Therefore, Singularity (Kurtzer et al., 2017) is generally more widespread on HPC clusters as it was specifically developed for it.

With the capacity to provide computational resources, web-based platforms, such as DAnIEL (Loos et al., 2021) and SCATA may be simply used through a web browser on any operating system. Additionally, some other pipelines, such as FROGS and LotuS2, can also be accessed through Galaxy websites, and nf-core/ampliseq (Straub et al., 2020) can be launched via Nextflow Tower (a monitoring and management platform for Nextflow workflows).

The increasing complexity of bioinformatics pipelines, which consist of a large number of computational steps, encouraged the development of workflow management systems capable of orchestrating in a scalable and reproducible manner (Mölder et al., 2021; Wratten et al., 2021). Workflow managers allow pipelines to resume after a failure and start from the last successfully completed step, automate pipeline execution triggered by input or reference data updates, and perform parameter exploration. Nextflow (Di Tommaso et al., 2017) and Snakemake (Koster & Rahmann, 2012; Mölder et al., 2021) are among the most prominent workflow management systems in the field of bioinformatics. They simplify pipeline development, maximize resource usage efficiency, and handle installation and versioning of the software dependencies (e.g., using Docker and Singularity containers or conda environments). These systems allow running workflow steps in parallel locally or using resources of HPC clusters or commercial cloud computing providers (Amazon web services (Bai et al., 2019), Microsoft Azure (Copeland et al., 2015), Google Cloud (Hussain & Aleem, 2018)) almost without the need to adapt pipeline code to a specific platform architecture. MetaWorks, dadasnake, Tourmaline (Thompson et al., 2022), and Cascabel (Asbun et al., 2020) are examples of Snakemake-based pipelines, while nf-core/ampliseq, eDNAflow and NextITS were developed using Nextflow.

The interface

Generally, the Unix-based command-line interfaces (CLI; commands are typed into a terminal) are often preferred by analysts with bioinformatics experience. That is because most of the pipelines are developed as CLI-runnable software that can be operated on HPC clusters, but also due to the flexibility and availability of applying various custom processes to manage the data effectively. Although the CLI tools offer numerous advantages, using a CLI might be intimidating for users with less programming experience. To facilitate the analysis of metabarcoding data by non-bioinformaticians, APSCALE, CoMA (Hupfauf et al., 2020), gDAT (Vasar et al., 2021), PipeCraft2 and SEED 2 (Vetrovský et al., 2018) provide a graphical user interface (GUI; interaction via clickable graphical icons; Table 1, Fig. 2) as a front-end for specifying the settings of the bioinformatics analyses, which will be executed on the back-end. Depending on the architecture, the GUI-based applications may require more RAM than CLI pipelines. Pipelines that have web server support (DAnIEL, SCATA) or have been implemented into Galaxy server (LotuS2, FROGS, QIIME 2) naturally possess a web-based GUI for specifying the settings of the analysis. Some software that is wrapped into GUI may also be executed through CLI (Table 1).

Marker-specific pipelines

A marker (i.e., 'DNA barcode') is a taxonomically informative gene fragment that is utilized for species identifications in the presence of reference sequence data. Bioinformatics processes combined in a pipeline may be specifically designed to analyze amplicons from a specific marker, i.e., the analytical steps may depend on the characteristics of the amplicons. For example, when processing ITS amplicon data, it is common to remove conservative flanking genes of ITS for accurate taxonomic classification purposes (Vu et al., 2022; Tedersoo et al., 2022). When processing sequences from the COI gene, removing the co-amplified putative nuclear mitochondrial pseudogenes (NUMTs) is highly recommended (Song et al., 2008; Porter & Hajibabaei., 2021; Greedy et al., 2021). The subsections below outline the herein-considered marker-specific and multi-marker pipelines and highlight some of the results from their benchmarking trials.

Prokaryotic 16S rRNA

Amplicon sequencing targeting the 16S rRNA gene is commonly utilized to investigate microbiomes from various ecosystems/substrates (Knight et al., 2018; Pollock et al., 2018; Staats et al., 2016). The 16S gene sequence is roughly 1,500 bp in length and contains nine distinct hypervariable regions (V1–V9). The V4 hypervariable region is most often used in short-read sequencing, whereas full-length 16S analyses are becoming increasingly utilized with the increased quality, availability, and decreasing costs of long-read sequencing methods. For processing 16S amplicons, mothur, USEARCH, QIIME 2 and DADA2 are the most used ones. Recently established pipelines such as dadaist2, dadasnake, nf-core/ampliseq, Tourmaline also wrap QIIME 2 and/or DADA2 functionalities and are thus optimized for 16S (but not exclusively) analyses. BIOCOM-PIPE, Cascabel, CoMA, LotuS2, MICCA, PEMA and FROGS have also benchmarked their pipelines using 16S data sets. However, since the bioinformatics processing of 16S amplicon data was at the forefront of metabarcoding data analyses before the wide-scale utilization of other markers, other multi-marker pipelines (Table 1) that consist of critical filtering steps may also be used to process 16S reads.

Testing different workflows on 16S V4 amplicon mock data (known composition of taxa in a sample), Straub et al. (2020) found QIIME 2 pipeline with the DADA2 plugin being the most optimal compared to mothur, QIIME 1, and MEGAN (Huson et al., 2007) workflows. Based on the benchmarking results, the nf-core/ampliseq pipeline was developed which demonstrated a high degree of similarity with the results produced by QIIME 2. Prodan et al. (2020) reported good performance of all tested ASV workflows (DADA2, QIIME 2 deblur, UNOISE3), but with slight variations in their sensitivity and specificity to detect mock community members. In the latter study, two OTU workflows also performed well (UPARSE, mothur; but not QIIME 1-uclust), but with lower specificity than ASV pipelines. A more recent study by Özkurt et al. (2022) reported the higher accuracy of LotuS2 compared with QIIME 2, DADA2, and PipeCraft2. The LotuS2 pipeline runs with stringent read filtering and implements a unique feature, a 'seed extension' algorithm, that improves the quality of a feature's representative sequence. By introducing the CoMA pipeline that uses LotuS1/2 (Hildebrand, et al., 2014) at its core, Hupfauf et al. (2020) reported a good performance of all tested pipelines (CoMA, QIIME 2, mothur). However, some degree of variability was evidently

depending on the test dataset. In general, the lack of consensus as to the 'best performing pipeline' illustrates the importance of the underlying dataset properties. Considering the dataset's characteristics under the operation, tweaking, and fine-tuning the settings of different pipelines may further, at least to some extent, diminish the variability in their accuracy.

ITS rRNA

The nuclear ribosomal internal transcribed spacer (ITS) region is a standard marker in fungal metabarcoding studies (Nilsson et al., 2019). It is also taxonomically informative in other eukaryotic groups (e.g., flowering plants, mites, springtails; Banchi et al., 2020; Ben-David et al., 2007; Anslan & Tedersoo., 2015). The ITS region is highly-variable in length among eukaryotic groups, complicating the bioinformatics analysis steps that rely on aligning (such as e.g., mothur OTU clustering) or require uniform sequence length (such as e.g., deblur). Pipelines such as NextITS, PIPITS (Gweon et al., 2015), and DAnIEL are developed explicitly for ITS amplicon analyses. Those pipelines implement the extraction of ITS sub-regions (ITS1/ITS2, or full ITS) to exclude flanking conservative regions (18S/5.8S/28S), which is optimal for taxonomic assignment accuracy (Vu et al., 2022, Bengtsson-Palme et al., 2013). SCATA is also optimized for the ITS region, and for other amplicon sequences which cannot be easily aligned. However, a few other universal pipelines, such as LotuS2, SEED 2, nf-core/ampliseq, PipeCraft2, MetaWorks, dadasnake, FROGS (all using ITSx; Bengtsson-Palme et al. 2013), and QIIME 2 (using the ITSxpress plugin; Rivers et al., 2018) incorporate the step for extracting the ITS sub-regions for optimal processing of ITS amplicon data. Because the ITS-subregions of some fungal groups may not sufficiently overlap during the paired-end data assembly process, FROGS, PipeCraft2, Dadaist2 and Cascabel (latter two without ITSx) implement settings to also include non-assembled reads to ensure that taxa with longer ITS regions are not excluded (Bernard et al., 2021).

Although AMPtk, DADA2, eDNAflow, and gDAT were validated using ITS reads, these pipelines lack a step to clip the flanking regions from ITS reads. While ITS extraction tools may eliminate some fungal strains from the data, many false-positive molecular units are generated when this extraction process is excluded (Pauvert et al., 2019). To mitigate the detection of false-negatives, the exclusion of the ITS extraction may be more appropriate if the aim is to find specific target taxa, whereas the ITS extraction operation should be included in community ecology studies (Pauvert et al., 2019).

Tested on technical replicates from soil samples (i.e. DNA from the same sample sequenced twice), compositional matrices of ITS data from QIIME 2 and LotuS2 were more reproducible than native DADA2, where the latter did not incorporate an ITS extraction step (Özkurt et al., 2022). Differences in the ITS amplicon data analyses among various software (PipeCraft1, QIIME 2, PIPITS, LotuS1, and custom pipeline compiled on Galaxy platform) were evident also in the study by Anslan et al. (2018) where QIIME 2 and Galaxy-based pipelines did not include the ITS extraction step (because it was not yet implemented). Although the inclusion of ITS region extraction step lowers the amount of non-target features, the latter study concluded that none of the tested workflows were able to fully filter out the erroneous sequences, which contributed to the demonstrated differences between pipelines.

COI

Found in the mitochondria, the cytochrome oxidase subunit I (COI/CO1/cox1) is a standard animal barcode (Hebert et al., 2003; Hajibabaei et al., 2011). Compared with other suitable markers (e.g., mt 16S, ITS, 28S) for most metazoan groups, the reference database of the COI is vast (Porter & Hajibabaei, 2018) and COI fragments are extensively used in metabarcoding studies.

Metabarcoding of metazoan communities is increasingly employed in ecology, but the strategies for analyzing the sequencing data vary largely across studies. Generally, the metabarcoding studies utilizing protein-coding genes (such as COI) have largely followed the bioinformatic workflows designed to characterize microbial diversity without adapting the workflows to the characteristics of protein-coding markers (Creedy et al., 2021). When processing protein-coding markers, the noise of nuclear mitochondrial pseudogenes (NUMTs) may inflate the richness estimates and thus introduce biases in biodiversity research using metabarcoding (Porter & Hajibabaei, 2021). Thus, the amino acid translation, but also the length of the read should be used to identify erroneous sequences (Creedy et al., 2021). Of the pipelines reviewed here, MetaWorks and VTAM implement a step of removing putative NUMTs, which alleviates the burden of manual curation of the features to produce more accurate richness estimates. The multi-marker amplicon processing platform PipeCraft2 has also wrapped MetaWorks strategy of the pseudogene removal step. Apart from the full pipelines, the multi-sample features matrix may be processed with metaMATE (Andújar et al., 2021) to remove putative NUMTs and other erroneous sequences (based on e.g., length and relative read abundance). Additionally, DARN (Zafeiropoulos et al., 2021), which makes use of the phylogenetic tree, aids in filtering out non-target features and upon denoising, the characteristics of protein coding genes are also accounted for in the DnoisE (Antich et al., 2021). We will most likely see the latter module integrated into the already established pipelines in the near future.

Other markers and multi-marker pipelines

Besides the above-mentioned markers, other popular markers used for metabarcoding are mt 16S rRNA for Metazoa, mt 12S rRNA for fish (Miya et al., 2020), 18S rRNA for protists and other eukaryotes, 28S rRNA for nematodes and eukaryotes in general, rbcL for diatoms (Rimet et al., 2019), rbcL+matK and trnL for plants (CBOL Plant Working Group et al., 2009; Taberlet et al., 2007), and 23S rRNA for photosynthetic microbes (Djemiel et al., 2020). A variety of pipelines have been applied for the analyses of the amplicon sequences from these markers. For example, MICCA, DADA2 for 18S rRNA (Harrison et al., 2021; Minerovic et al., 2020); DADA2, OBITools for mt 16S (Thomsen & Sigsgaard, 2019; Marquina et al., 2019); and custom built pipelines (using multiple third-party sequence data analysis tools) for other markers above (Westfall et al., 2019; Liu & Zhang, 2021; Elbrecht et al., 2016; Anslan et al., 2021). Benchmarked on mt 12S reads from both simulated and real eDNA data, the Barque pipeline demonstrated a small sensitivity improvement over QIIME 2 and OBITools (Mathon et al., 2021). Moreover, another VSEARCH-based custom pipeline found in the latter study, which was designed to match Barque's performance by adjusting the parameters and threshold, showed the same mean sensitivity as Barque, demonstrating that the careful choice of the tools for the required task provides accurate results.

Table 1 lists multi-marker software that may be utilized for various markers. All of the developed application software contain the most crucial steps for basic metabarcoding data analyses, but the suitability of a software or workflow steps for a given marker should be assessed. For example, considering the length variability and align ability of the amplicon set is important when some pipeline steps (e.g., clustering) use alignment-based methods (such as in mothur) or require uniform read lengths (such as deblur denoising). When working e.g., rbcL amplicons (or amplicons from any other protein coding gene), validation is needed to ensure that the generated features do not represent potential pseudogenes (or off-target taxa) for biodiversity analyses. Some multi-marker pipelines incorporate marker-specific steps, e.g., extracting the ITS region, removing putative pseudogenes and off-target features (Supplementary File 1). Using a pipeline that is not restricted to a certain marker gene, but where the above listed automated filtering processes are lacking, a manual feature curation step is usually required to filter out bioinformatically unfiltered noise or to validate that most of the noise has already been removed. Depending on the study context, different analytical pipelines may yield highly compatible results (e.g., Kang et al., 2021; Baltrušiset et al., 2022), but the outcome and interpretation may also vary considerably (Anslan et al., 2018; Pauvert et al., 2019; Straub et al., 2020; Bailet et al., 2020) without the validation of the software suitability for a given marker.

Concluding remarks

The development of a wide range of metabarcoding data analysis pipelines illustrates the need for 'easy-to-use' software, but also of specific customized workflows depending on the underlying sequencing data set. Although most of the pre-compiled pipelines largely mirror the functionalities of several software suites by incorporating steps from algorithms providing software suites, they offer easily executable automated alternatives for users with less bioinformatics experience. Additionally, many pre-compiled pipelines are supplemented with several possibilities for downstream analyses by wrapping various third-party tools. Applying different workflows on the same data will always demonstrate a certain level of variation among pipelines. These variations are usually most obvious in terms of the reported number of features. This generally derives from variations in filtering out spurious and low-abundant sequences (e.g. Edgar, 2017; Prodan et al., 2020). Therefore, one pipeline may produce a higher number of features per sample and the other much less, but the correlations between sample-wise richness from one to another result are in most cases very high (Kang et al., 2021; Baltrušis et al., 2022). However, depending on the analyzed data set, this correlation pattern may be the opposite (Nearing et al., 2018) and pipeline settings should be carefully considered, especially when identifying rare taxa is imperative. Thus, although the automated pipelines have made the analyses easier and more reproducible, expertise is still required to validate the accuracy of the biological results. It is noteworthy that a pipeline's performance measured on mock community samples with relatively few species may vary when applied to a complex data set originating from environmental samples. Nevertheless, including a mock community control sample(s) in a study will certainly aid in identifying false positives and false negatives. A robust sense of the community patterns may be obtained by applying 'default' parameter values but

fine-tuning of the parameters may be required to find an appropriate compromise between false positive removal and retention of true detections.

Table 1 and Figure 2 are aiming to provide assistance in narrowing down the desirable pipelines for the task. Once the potential target workhorses have been selected, one would naturally need to explore the respective user guides for more detailed information about the underlying procedures.

Acknowledgements

This work was supported by the European Regional Development Fund and the programme Mobilitas Pluss (MOBTP198). MH and TMP received funding from Genome Canada and Ontario Genomics through the Sequencing the Rivers for Environmental Assessment and Monitoring (STREAM) project. DS acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy, cluster of Excellence EXC2124 "Controlling microbes to fight infection" (CMFI), project ID 390838134. We wish to offer our heartfelt thanks to Sébastien Terrat, the leading developer of BIOCOM-PIPE (including ReClustOR) for the support extended to certain points in this tool. TV was supported by the Czech Science Foundation (21-17749S). We thank other FROGS'members Vincent Darbot, Lucas Auer and Olivier Rué, and Frédéric Mahé (swarm software author) for their fruitful exchanges on the ASV/OTU/cluster terminology. EC received funding through an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103449.

References

Albanese, D., Fontana, P., De Filippo, C., Cavalieri, D., & Donati, C. (2015). MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. Scientific Reports, 5(1), 1–7. https://doi.org/10.1038/srep09743

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17), 3389-3402. https://doi.org/10.1093/nar/25.17.3389

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A. J., Vogler, A. P., & Emerson, B. C. (2021). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. Molecular Ecology Resources, 21(6), 1772–1787. https://doi.org/10.1111/1755-0998.13337

Anslan, S., & Tedersoo, L. (2015). Performance of cytochrome c oxidase subunit I (COI), ribosomal DNA Large Subunit (LSU) and Internal Transcribed Spacer 2 (ITS2) in DNA barcoding of Collembola. European Journal of Soil Biology, 69, 1-7. https://doi.org/10.1016/j.ejsobi.2015.04.001

Anslan, S., Bahram, M., Hiiesalu, I., & Tedersoo, L. (2017). PipeCraft: Flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. Molecular Ecology Resources, 17(6), e234–e240. https://doi.org/10.1111/1755-0998.12692

Anslan, S., Mikryukov, V., Armolaitis, K., Ankuda, J., Lazdina, D., Makovskis, K., ... & Tedersoo, L. (2021). Highly comparable metabarcoding results from MGI-Tech and Illumina sequencing platforms. PeerJ, 9, e12254. https://doi.org/10.7717/peerj.12254

Anslan, S., Nilsson, R. H., Wurzbacher, C., Baldrian, P., Leho Tedersoo, & Bahram, M. (2018). Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. MycoKeys, (39), 29–40. https://doi.org/10.3897/mycokeys.39.28109

Ansorge, R., Birolo, G., James, S. A., & Telatin, A. (2021). Dadaist2: a toolkit to automate and simplify statistical analysis and plotting of metabarcoding experiments. International journal of molecular sciences, 22(10), 5309. https://doi.org/10.3390/ijms22105309

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. BMC Bioinformatics, 22(1), 177. https://doi.org/10.1186/s12859-021-04115-6.

Antich, A., Palacín, C., Turon, X., & Wangensteen, O. S. (2022). DnoisE: distance denoising by entropy. An open-source parallelizable alternative for denoising sequence datasets. PeerJ, 10, e12758. https://doi.org/10.7717/peerj.12758

Asbun, A. A., Besseling, M. A., Balzano, S., van Bleijswijk, J. D. L., Witte, H. J., Villanueva, L., & Engelmann, J. C. (2020). Cascabel: A Scalable and Versatile Amplicon Sequence Data Analysis Pipeline Delivering Reproducible and Documented Results. In Frontiers in Genetics (Vol. 11). https://doi.org/10.3389/fgene.2020.489357

Bai, J., Jhaney, I., & Wells, J. (2019). Developing a reproducible microbiome data analysis pipeline using the Amazon web services cloud for a cancer research group: proof-of-concept study. JMIR medical informatics, 7(4), e14667. https://doi.org/10.2196/14667

Bailet, B., Apothéloz-Perret-Gentil, L., Baričević, A., Chonova, T., Franc, A., Frigerio, J. M., ... & Kahlert, M. (2020). Diatom DNA metabarcoding for ecological assessment: Comparison among bioinformatics pipelines used in six European countries reveals the need for standardization. Science of the Total Environment, 745, 140948. https://doi.org/10.1016/j.scitotenv.2020.140948

Baloğlu, B., Chen, Z., Elbrecht, V., Braukmann, T., MacDonald, S., & Steinke, D. (2021). A workflow for accurate metabarcoding using nanopore MinION sequencing. Methods in Ecology and Evolution, 12(5), 794-804. https://doi.org/10.1111/2041-210X.13561

Baltrušis, P., Halvarsson, P., & Höglund, J. (2022). Estimation of the impact of three different bioinformatic pipelines on sheep nemabiome analysis. Parasites & Vectors, 15(1), 1-12. https://doi.org/10.1186/s13071-022-05399-0

Banchi, E., Ametrano, C. G., Greco, S., Stanković, D., Muggia, L., & Pallavicini, A. (2020). PLANiTS: A curated sequence reference dataset for plant ITS DNA metabarcoding. Database, 2020(baz155). https://doi.org/10.1093/database/baz155

Ben-David, T., Melamed, S., Gerson, U., & Morin, S. (2007). ITS2 sequences as barcodes for

identifying and analyzing spider mites (Acari: Tetranychidae). Experimental and Applied Acarology, 41(3), 169-181. https://doi.org/10.1186/s13071-022-05399-0

Bengtsson-Palme, J., Ryberg, M., Hartmann, M., Branco, S., Wang, Z., Godhe, A., ... & Nilsson, R. H. (2013). Improved software detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of fungi and other eukaryotes for analysis of environmental sequencing data. Methods Ecol Evol. 2013; 4 (10): 914–9. https://doi.org/10.1111/2041-210X.12073

Bernard, M., Rué, O., Mariadassou, M., & Pascal, G. (2021). FROGS: a powerful tool to analyse the diversity of fungi with special management of internal transcribed spacers. Briefings in Bioinformatics, 22(6). https://doi.org/10.1093/bib/bbab318

Bokulich, N. A., Kaehler, B. D., Rideout, J. R., Dillon, M., Bolyen, E., Knight, R., ... & Gregory Caporaso, J. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome, 6(1), 1-17. https://doi.org/10.1186/s40168-018-0470-z

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology, 37(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: a unix-inspired software package for DNA metabarcoding. Molecular Ecology Resources, 16(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. Molecular Ecology Resources, 21(6), 1904-1921. https://doi.org/10.1111/1755-0998.13398

Brown, S. P., Veach, A. M., Rigdon-Huss, A. R., Grond, K., Lickteig, S. K., Lothamer, K., ... & Jumpponen, A. (2015). Scraping the bottom of the barrel: are rare high throughput sequences artifacts?. fungal ecology, 13, 221-225.

Bruce, K., Blackman, R. C., Bourlat, S. J., Hellström, M., Bakker, J., Bista, I., ... & Deiner, K. (2021). A practical guide to DNA-based methods for biodiversity assessment. https://doi.org/10.3897/ab.e68634

Buchner, D., Macher, T.-H., & Leese, F. (2022). APSCALE: advanced pipeline for simple yet comprehensive analyses of DNA metabarcoding data. Bioinformatics , 38(20), 4817–4819. https://doi.org/10.1093/bioinformatics/btac588

Callahan, B. J., Grinevich, D., Thakur, S., Balamotis, M. A., & Yehezkel, T. B. (2021). Ultra-accurate microbial amplicon sequencing with synthetic long reads. Microbiome, 9(1), 130. https://doi.org/10.1186/s40168-021-01072-3

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal, 11(12), 2639-2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7), 581–583. https://doi.org/10.1038/nmeth.3869

Callahan, B. J., Wong, J., Heiner, C., Oh, S., Theriot, C. M., Gulati, A. S., ... & Dougherty, M. K. (2019). High-throughput amplicon sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. Nucleic acids research, 47(18), e103. https://doi.org/10.1093/nar/gkz569

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., ... & Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature methods, 7(5), 335-336. https://doi.org/10.1038/nmeth.f.303

Carlsen, T., Aas, A. B., Lindner, D., Vrålstad, T., Schumacher, T., & Kauserud, H. (2012). Don't make a mista (g) ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies?. Fungal Ecology, 5(6), 747-749.

Carøe, C., & Bohmann, K. (2020). Tagsteady: a metabarcoding library preparation protocol to avoid false assignment of sequences to samples. Molecular Ecology Resources, 20(6), 1620-1631. https://doi.org/10.1111/1755-0998.13227

Castaño, C., Berlin, A., Brandström Durling, M., Ihrmark, K., Lindahl, B. D., Stenlid, J., ... & Olson, Å. (2020). Optimized metabarcoding with Pacific Biosciences enables semi-quantitative analysis of fungal communities. New Phytologist, 228(3). https://doi.org/10.1111/nph.16731

CBOL Plant Working Group 1, Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., ... & Little, D. P. (2009). A DNA barcode for land plants. Proceedings of the National Academy of Sciences, 106(31), 12794-12797. https://doi.org/10.1073/pnas.0905845106

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics, 34(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Community, G. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic acids research, 50(W1), W345–W351. https://doi.org/10.1093/nar/gkac247

Compson, Z. G., McClenaghan, B., Singer, G. A., Fahner, N. A., & Hajibabaei, M. (2020). Metabarcoding from microbes to mammals: comprehensive bioassessment on a global scale. Frontiers in Ecology and Evolution, 8, 581835. https://doi.org/10.3389/fevo.2020.581835

Copeland, M., Soh, J., Puca, A., Manning, M., & Gollob, D. (2015). Microsoft azure. New York, NY, USA:: Apress, 3-26.

Couton, M., Baud, A., Daguin-Thiébaut, C., Corre, E., Comtet, T., & Viard, F. (2021). High-throughput sequencing on preservative ethanol is effective at jointly examining infraspecific and taxonomic diversity, although bioinformatics pipelines do not perform equally. Ecology and evolution, 11(10), 5533-5546. https://doi.org/10.1002/ece3.7453

Creedy, T. J., Andujar, C., Meramveliotakis, E., Noguerales, V., Overcast, I., Papadopoulou, A., ... & Arribas, P. (2022). Coming of age for COI metabarcoding of whole organism community DNA: towards bioinformatic harmonisation. Molecular Ecology Resources, 22(3), 847-861. https://doi.org/10.1111/1755-0998.13502

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., ... & Meyer, R. S. (2019). Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. Methods in Ecology and Evolution, 10(9), 1469-1475. https://doi.org/10.1111/2041-210X.13214

De Santiago, A., Pereira, T. J., Mincks, S. L., & Bik, H. M. (2022). Dataset complexity impacts both MOTU delimitation and biodiversity estimates in eukaryotic 18S rRNA metabarcoding studies. Environmental DNA, 4(2), 363-384. https://doi.org/10.1002/edn3.255

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. Nature Biotechnology, 35(4), 316–319. https://doi.org/10.1038/nbt.3820

Djemiel, C., Dequiedt, S., Karimi, B., Cottin, A., Girier, T., El Djoudi, Y., Wincker, P., Lelièvre, M., Mondy, S., Chemidlin Prévost-Bouré, N., Maron, P.-A., Ranjard, L., & Terrat, S. (2020). BIOCOM-PIPE: a new user-friendly metabarcoding pipeline for the characterization of microbial diversity from 16S, 18S and 23S rRNA gene amplicons. BMC Bioinformatics, 21(1), 492. https://doi.org/10.1186/s12859-020-03829-3

Djemiel, C., Plassard, D., Terrat, S., Crouzet, O., Sauze, J., Mondy, S., ... & Maron, P. A. (2020). µgreen-db: a reference database for the 23S rRNA gene of eukaryotic plastids and cyanobacteria. Scientific reports, 10(1), 1-11. https://doi.org/10.1038/s41598-020-62555-1

Durling, M. B., Clemmensen, K. E., Stenlid, J., & Lindahl, B. (2011). SCATA-An efficient bioinformatic pipeline for species identification and quantification after high-throughput sequencing of tagged amplicons.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26(19), 2460-2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2016). SINTAX: a simple non-Bayesian taxonomy classifier for 16S and ITS sequences. biorxiv, 074161. https://doi.org/10.1101/074161

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv, 081257. https://doi.org/10.1101/081257

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed-and open-reference OTUs. PeerJ, 5, e3889. https://doi.org/10.7717/peerj.3889

Edgar, R. C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS

sequences. PeerJ, 6, e4652. https://doi.org/10.7717/peerj.4652

Edgar, R. C. (2018). UNCROSS2: identification of cross-talk in 16S rRNA OTU tables. BioRxiv, 400762. https://doi.org/10.1101/400762

Edgar, R. C., & Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics, 31(21), 3476-3482. https://doi.org/10.1093/bioinformatics/btv401

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics, 27(16), 2194-2200. https://doi.org/10.1093/bioinformatics/btr381

Elbrecht, V., Taberlet, P., Dejean, T., Valentini, A., Usseglio-Polatera, P., Beisel, J. N., ... & Leese, F. (2016). Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. PeerJ, 4, e1966. https://doi.org/10.7717/peerj.1966

Escudié F, Auer L, Bernard M, Mariadassou M, Cauquil L, Vidal K, Maman S, Hernandez-Raquet G, Combes S, Pascal G. FROGS: Find, Rapidly, OTUs with Galaxy Solution. Bioinformatics. 2018 Apr 15;34(8):1287-1294. https://doi.org/10.1093/bioinformatics/btx791

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature communications, 8(1), 1-11. https://doi.org/10.1038/s41467-017-01312-x

Furneaux, B., Bahram, M., Rosling, A., Yorou, N. S., & Ryberg, M. (2021). Long-and short-read metabarcoding technologies reveal similar spatiotemporal structures in fungal communities. Molecular Ecology Resources, 21(6), 1833-1849. https://doi.org/10.1111/1755-0998.13387

Glassman, S. I., & Martiny, J. B. (2018). Broadscale ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. MSphere, 3(4), e00148-18. https://doi.org/10.1128/mSphere.00148-18

Gold, Z., Curd, E. E., Goodwin, K. D., Choi, E. S., Frable, B. W., Thompson, A. R., ... & Barber, P. H. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. Molecular ecology resources, 21(7), 2546-2564. https://doi.org/10.1111/1755-0998.13450

González, A., Dubut, V., Corse, E., Mekdad, R., Dechatre, T., Castet, U., ... & Meglécz, E. (2023). VTAM: A robust pipeline for validating metabarcoding data using controls. Computational and Structural Biotechnology Journal. https://doi.org/10.1016/j.csbj.2023.01.034

Gweon, H. S., Oliver, A., Taylor, J., Booth, T., Gibbs, M., Read, D. S., Griffiths, R. I., & Schonrogge, K. (2015). PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. Methods in Ecology and Evolution / British Ecological Society, 6(8), 973–980. https://doi.org/10.1111/2041-210X.12399

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A., & Baird, D. J. (2011). Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS one, 6(4), e17497. https://doi.org/10.1371/journal.pone.0017497

Harrison, J. P., Chronopoulou, P. M., Salonen, I. S., Jilbert, T., & Koho, K. A. (2021). 16S and 18S rRNA gene metabarcoding provide congruent information on the responses of sediment communities to eutrophication. Frontiers in Marine Science, 8, 708716. https://doi.org/10.3389/fmars.2021.708716

Hebert, Paul DN, et al. "Biological identifications through DNA barcodes." Proceedings of the Royal Society of London. Series B: Biological Sciences 270.1512 (2003): 313-321. https://doi.org/10.1098/rspb.2002.2218

Heeger, F., Bourne, E. C., Baschien, C., Yurkov, A., Bunk, B., Spröer, C., ... & Monaghan, M. T. (2018). Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. Molecular Ecology Resources, 18(6), 1500-1514. https://doi.org/10.1111/1755-0998.12937

Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P., & Raes, J. (2014). LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome, 2(1), 1-7. https://doi.org/10.1186/2049-2618-2-30

Hleap, J. S., Littlefair, J. E., Steinke, D., Hebert, P. D., & Cristescu, M. E. (2021). Assessment of current taxonomic assignment strategies for metabarcoding eukaryotes. Molecular Ecology Resources, 21(7), 2190-2203. https://doi.org/10.1111/1755-0998.13407

Hupfauf, S., Etemadi, M., Juárez, M. F.-D., Gómez-Brandón, M., Insam, H., & Podmirseg, S. M. (2020). CoMA – an intuitive and user-friendly pipeline for amplicon-sequencing data analysis. In PLOS ONE (Vol. 15, Issue 12, p. e0243241). https://doi.org/10.1371/journal.pone.0243241

Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. Environmental microbiology, 12(7), 1889-1898. https://doi.org/10.1111/j.1462-2920.2010.02193.x

Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. Genome Research, 17(3), 377–386. https://doi.org/10.1101/gr.5969107

Hussain, A., & Aleem, M. (2018). GoCJ: Google cloud jobs dataset for distributed and cloud computing infrastructures. Data, 3(4), 38. https://doi.org/10.3390/data3040038

Kaehler, B. D., Bokulich, N. A., McDonald, D., Knight, R., Caporaso, J. G., & Huttley, G. A. (2019). Species abundance information improves sequence taxonomy classification accuracy. Nature communications, 10(1), 4643. https://doi.org/10.1038/s41467-019-12669-6

Kang, W., Anslan, S., Börner, N., Schwarz, A., Schmidt, R., Künzel, S., ... & Schwalb, A. (2021). Diatom metabarcoding and microscopic analyses from sediment samples at Lake Nam Co, Tibet: The effect of sample-size and bioinformatics on the identified communities. Ecological Indicators, 121, 107070. https://doi.org/10.1016/j.ecolind.2020.107070

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. Nature Reviews. Microbiology, 16(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Koster, J., & Rahmann, S. (2012). Snakemake--a scalable bioinformatics workflow engine. In Bioinformatics (Vol. 28, Issue 19, pp. 2520–2522). https://doi.org/10.1093/bioinformatics/bts480

Kurtzer, G. M., Sochat, V., & Bauer, M. W. (2017). Singularity: Scientific containers for mobility of compute. PloS one, 12(5), e0177459. https://doi.org/10.1371/journal.pone.0177459

Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. Briefings in bioinformatics, 17(1), 154-179. https://doi.org/10.1093/bib/bbv029

Lear, G., Dickie, I., Banks, J., Boyer, S., Buckley, H. L., Buckley, T. R., ... & Holdaway, R. (2018). Methods for the extraction, storage, amplification and sequencing of DNA from environmental samples. New Zealand Journal of Ecology, 42(1), 10-50A. https://doi.org/10.20417/nzjecol.42.9

Lindgreen, S. (2012). AdapterRemoval: easy cleaning of next-generation sequencing reads. BMC research notes, 5(1), 1-7. https://doi.org/10.1186/1756-0500-5-337

Liu, J., & Zhang, H. (2021). Combining multiple markers in environmental DNA metabarcoding to assess deep-sea benthic biodiversity. Frontiers in Marine Science, 8, 684955. https://doi.org/10.3389/fmars.2021.684955

Loos, D., Zhang, L., Beemelmanns, C., Kurzai, O., & Panagiotou, G. (2021). DAnIEL: A User-Friendly Web Server for Fungal ITS Amplicon Sequencing Data. Frontiers in Microbiology, 12, 720513. https://doi.org/10.3389/fmicb.2021.720513
Mahé, F., Czech, L., Stamatakis, A., Quince, C., de Vargas, C., Dunthorn, M., & Rognes, T. (2022). Swarm v3: towards tera-scale amplicon clustering. Bioinformatics, 38(1), 267-269. https://doi.org/10.1093/bioinformatics/btab493

Marquina, D., Esparza-Salas, R., Roslin, T., & Ronquist, F. (2019). Establishing arthropod community composition using metabarcoding: Surprising inconsistencies between soil samples and preservative ethanol and homogenate from Malaise trap catches. Molecular ecology resources, 19(6), 1516-1530. https://doi.org/10.1111/1755-0998.13071

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), 10-12. https://doi.org/10.14806/ej.17.1.200

Mathon, L., Valentini, A., Guérin, P. E., Normandeau, E., Noel, C., Lionnet, C., ... & Manel, S. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. Molecular Ecology Resources, 21(7), 2565-2579. https://doi.org/10.1111/1755-0998.13430

McGee, K. M., Robinson, C. V., & Hajibabaei, M. (2019). Gaps in DNA-based biomonitoring across the globe. Frontiers in Ecology and Evolution, 7, 337. https://doi.org/10.3389/fevo.2019.00337

Mikryukov V., Anslan S., Tedersoo L. NextITS: a pipeline for metabarcoding fungi and other eukaryotes with full-length ITS sequenced with PacBio. https://github.com/vmikk/NextITS

Minerovic, A. D., Potapova, M. G., Sales, C. M., Price, J. R., & Enache, M. D. (2020). 18S-V9 DNA metabarcoding detects the effect of water-quality impairment on stream biofilm eukaryotic assemblages. Ecological Indicators, 113, 106225. https://doi.org/10.1016/j.ecolind.2020.106225

Miya, M., Gotoh, R. O., & Sado, T. (2020). MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. Fisheries Science, 86(6), 939-970. https://doi.org/10.1007/s12562-020-01461-x

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. F1000Research, 10, 33. https://doi.org/10.12688/f1000research.29032.2

Mousavi-Derazmahalleh, M., Stott, A., Lines, R., Peverley, G., Nester, G., Simpson, T., ... & Christophersen, C. T. (2021). eDNAFlow, an automated, reproducible and scalable workflow for analysis of environmental DNA sequences exploiting Nextflow and Singularity. Molecular Ecology Resources, 21(5), 1697-1704. https://doi.org/10.1111/1755-0998.13356

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. PeerJ, 6, e5364. https://doi.org/10.7717/peerj.5364

Nilsson, R. H., Anslan, S., Bahram, M., Wurzbacher, C., Baldrian, P., & Tedersoo, L. (2019). Mycobiome diversity: high-throughput sequencing and identification of fungi. Nature Reviews. Microbiology, 17(2), 95–109. https://doi.org/10.1038/s41579-018-0116-y

Nilsson, R. H., Wurzbacher, C., Bahram, M., Coimbra, V. R., Larsson, E., Tedersoo, L., ... & Abarenkov, K. (2016). Top 50 most wanted fungi. MycoKeys, (12), 29-40. https://doi.org/10.3897/mycokeys.12.7553

Özkurt, E., Fritscher, J., Soranzo, N., Ng, D. Y. K., Davey, R. P., Bahram, M., & Hildebrand, F. (2022). LotuS2: an ultrafast and highly accurate tool for amplicon sequencing analysis. Microbiome, 10(1), 176. https://doi.org/10.1186/s40168-022-01365-1

Palmer, J. M., Jusino, M. A., Banik, M. T., & Lindner, D. L. (2018). Non-biological synthetic spike-in controls and the AMPtk software pipeline improve mycobiome data. PeerJ, 6, e4925. https://doi.org/10.7717/peerj.4925

Pauvert, C., Buee, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., ... & Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is

highly dependent on the metabarcoding pipeline. Fungal Ecology, 41, 23-33. https://doi.org/10.1016/j.funeco.2019.03.005

Plummer, E., Twin, J., Bulach, D. M., Garland, S. M., & Tabrizi, S. N. (2015). A comparison of three bioinformatics pipelines for the analysis of preterm gut microbiota using 16S rRNA gene sequencing data. Journal of Proteomics & Bioinformatics, 8(12), 283-291. https://doi.org/10.3389/fmicb.2020.01262

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. Applied and Environmental Microbiology, 84(7). https://doi.org/10.1128/AEM.02627-17

Porter, T. M., & Hajibabaei, M. (2018). Automated high throughput animal CO1 metabarcode classification. Scientific Reports, 8(1), 4226. https://doi.org/10.1038/s41598-018-22505-4

Porter, T. M., & Hajibabaei, M. (2020). Putting COI metabarcoding in context: The utility of exact sequence variants (ESVs) in biodiversity analysis. Frontiers in Ecology and Evolution, 8, 248. https://doi.org/10.3389/fevo.2020.00248

Porter, T. M., & Hajibabaei, M. (2021). Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. BMC bioinformatics, 22(1), 1-20. https://doi.org/10.1186/s12859-021-04180-x

Porter, T. M., & Hajibabaei, M. (2022). MetaWorks: A flexible, scalable bioinformatic pipeline for high-throughput multi-marker biodiversity assessments. PloS One, 17(9), e0274260. https://doi.org/10.1371/journal.pone.0274260

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. PLoS One, 15(1), e0227434. https://doi.org/10.1371/journal.pone.0227434

Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). Molecular ecology notes, 7(3), 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Reeder, J., & Knight, R. (2009). The'rare biosphere': a reality check. Nature methods, 6(9), 636-637. https://doi.org/10.1038/nmeth0909-636

Reitmeier, S., Hitch, T. C., Treichel, N., Fikas, N., Hausmann, B., Ramer-Tait, A. E., ... & Clavel, T. (2021). Handling of spurious sequences affects the outcome of high-throughput 16S rRNA gene amplicon profiling. ISME Communications, 1(1), 1-12. https://doi.org/10.1038/s43705-021-00033-z

Richardson, R. T., Bengtsson-Palme, J., & Johnson, R. M. (2017). Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Molecular Ecology Resources, 17(4), 760-769. https://doi.org/10.1111/1755-0998.12628

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., ... & Bouchez,

A. (2019). Diat. barcode, an open-access curated barcode library for diatoms. Scientific Reports, 9(1), 15116. https://doi.org/10.1038/s41598-019-51500-6

Rivers, A. R., Weber, K. C., Gardner, T. G., Liu, S., & Armstrong, S. D. (2018). ITSxpress: Software to rapidly trim internally transcribed spacer sequences with quality scores for marker gene analysis. F1000Research, 7. https://doi.org/10.12688/f1000research.15704.1

Rodriguez-Martinez, S., Klaminder, J., Morlock, M. A., Dalén, L., & Huang, D. T. (2022). The topological nature of tag jumping in environmental DNA metabarcoding studies. Molecular Ecology Resources. https://doi.org/10.1111/1755-0998.13745

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. PeerJ, 4, e2584. https://doi.org/10.7717/peerj.2584

Rosen, G. L., Reichenberger, E. R., & Rosenfeld, A. M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics, 27(1), 127-129. https://doi.org/10.1093/bioinformatics/btq619

Sato, M., Sugaya, N., Murakami, H., Imaizumi, A., Aburatani, S., Akutsu, T., & Horimoto, K. (2004). Remote homolog detection by match-node profile in hidden Markov model. In N. Callaos, K. Horimoto, J. Chen, & A. K. S. Chan (Eds.), 8th World Multi-Conference on Systemics, Cybernetics and Informatics, Vol Vii, Proceedings: Applications of Informatics and Cybernetics in Science and Engineering (pp. 27–34). Int Inst Informatics & Systemics. http://www.webofscience.com/wos/alldb/full-record/WOS:000227682900005

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., ... & Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology, 75(23), 7537-7541. https://doi.org/10.1128/AEM.01541-09

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. Molecular ecology resources, 15(6), 1289-1303. https://doi.org/10.1111/1755-0998.12402

Singer, G. A. C., Fahner, N. A., Barnes, J. G., McCarthy, A., & Hajibabaei, M. (2019). Comprehensive biodiversity analysis via ultra-deep patterned flow cell technology: a case study of eDNA metabarcoding seawater. Scientific reports, 9(1), 5991. https://doi.org/10.1038/s41598-019-42455-9

Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proceedings of the National Academy of Sciences of the United States of America, 105(36), 13486–13491. https://doi.org/10.1073/pnas.0803076105

Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T. W., & Kok, E. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. Analytical and Bioanalytical Chemistry, 408(17), 4615–4630. https://doi.org/10.1007/s00216-016-9595-8

Straub, D., Blackwell, N., Langarica-Fuentes, A., Peltzer, A., Nahnsen, S., & Kleindienst, S.

(2020). Interpretations of Environmental Microbial Community Studies Are Biased by the Selected 16S rRNA (Gene) Amplicon Sequencing Pipeline. Frontiers in Microbiology, 11, 550420. https://doi.org/10.3389/fmicb.2020.550420

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). Environmental DNA: For biodiversity research and monitoring. Oxford University Press. https://doi.org/10.1093/oso/9780198767220.001.0001

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental dna. Molecular ecology, 21(8), 1789-1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular ecology, 21(8), 2045-2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., ... & Willerslev, E. (2007). Power and limitations of the chloroplast trn L (UAA) intron for plant DNA barcoding. Nucleic acids research, 35(3), e14-e14. https://doi.org/10.1093/nar/gkl938

Tedersoo, L., & Anslan, S. (2019). Towards PacBio-based pan-eukaryote metabarcoding using full-length ITS sequences. Environmental Microbiology Reports, 11(5), 659-668. https://doi.org/10.1111/1758-2229.12776

Tedersoo, L., Albertsen, M., Anslan, S., & Callahan, B. (2021). Perspectives and benefits of high-throughput long-read sequencing in microbial ecology. Applied and environmental microbiology, 87(17), e00626-21. https://doi.org/10.1128/AEM.00626-21

Tedersoo, L., Bahram, M., Zinger, L., Nilsson, R. H., Kennedy, P. G., Yang, T., ... & Mikryukov, V. (2022). Best practices in metabarcoding of fungi: From experimental design to results. Molecular ecology, 31(10), 2769-2795. https://doi.org/10.1111/mec.16460

Terrat, S., Djemiel, C., Journay, C., Karimi, B., Dequiedt, S., Horrigue, W., ... & Ranjard, L. (2020). ReClustOR: a re-clustering tool using an open-reference method that improves operational taxonomic unit definition. Methods in Ecology and Evolution, 11(1), 168-180. https://doi.org/10.1111/2041-210X.13316

Thompson, L. R., Anderson, S. R., Den Uyl, P. A., Patin, N. V., Lim, S. J., Sanderson, G., & Goodwin, K. D. (2022). Tourmaline: A containerized workflow for rapid and iterable amplicon sequence analysis using QIIME 2 and Snakemake. *GigaScience*, *11*. giac066. https://doi.org/10.1093/gigascience/giac066

Thomsen, P. F., & Sigsgaard, E. E. (2019). Environmental DNA metabarcoding of wild flowers reveals diverse communities of terrestrial arthropods. Ecology and evolution, 9(4), 1665-1679. https://doi.org/10.1002/ece3.4809

Vasar, M., Davison, J., Neuenkamp, L., Sepp, S.-K., Young, J. P. W., Moora, M., & Öpik, M. (2021). User-friendly bioinformatics pipeline gDAT (graphical downstream analysis tool) for analysing rDNA sequences. Molecular Ecology Resources, 21(4), 1380–1392. https://doi.org/10.1111/1755-0998.13340

Vetrovský, T., Baldrian, P., & Morais, D. (2018). SEED 2: a user-friendly platform for amplicon high-throughput sequencing data analyses. Bioinformatics , 34(13), 2292–2294. https://doi.org/10.1093/bioinformatics/bty071

Vu, D., Nilsson, R. H., & Verkley, G. J. M. (2022). Dnabarcoder: An open-source software package for analysing and predicting DNA sequence similarity cutoffs for fungal sequence identification. Molecular Ecology Resources. https://doi.org/10.1111/1755-0998.13651

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and environmental microbiology, 73(16), 5261-5267. https://doi.org/10.1128/AEM.00062-07

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., ... & Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of the Total Environment, 678, 499-524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Westfall, K. M., Therriault, T. W., & Abbott, C. L. (2020). A new approach to molecular biosurveillance of invasive species using DNA metabarcoding. Global Change Biology, 26(2), 1012-1022. https://doi.org/10.1111/gcb.14886

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nature methods, 18(10), 1161-1168. https://doi.org/10.1038/s41592-021-01254-9

Zafeiropoulos, H., Gargan, L., Hintikka, S., Pavloudi, C., & Carlsson, J. (2021). The Dark mAtteR iNvestigator (DARN) tool: getting to know the known unknowns in COI amplicon data. Metabarcoding and Metagenomics, 5, e69657. https://doi.org/10.3897/mbmg.5.69657

Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., ... & Pafilis, E. (2020). PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. GigaScience, 9(3), giaa022. https://doi.org/10.1093/gigascience/giaa022

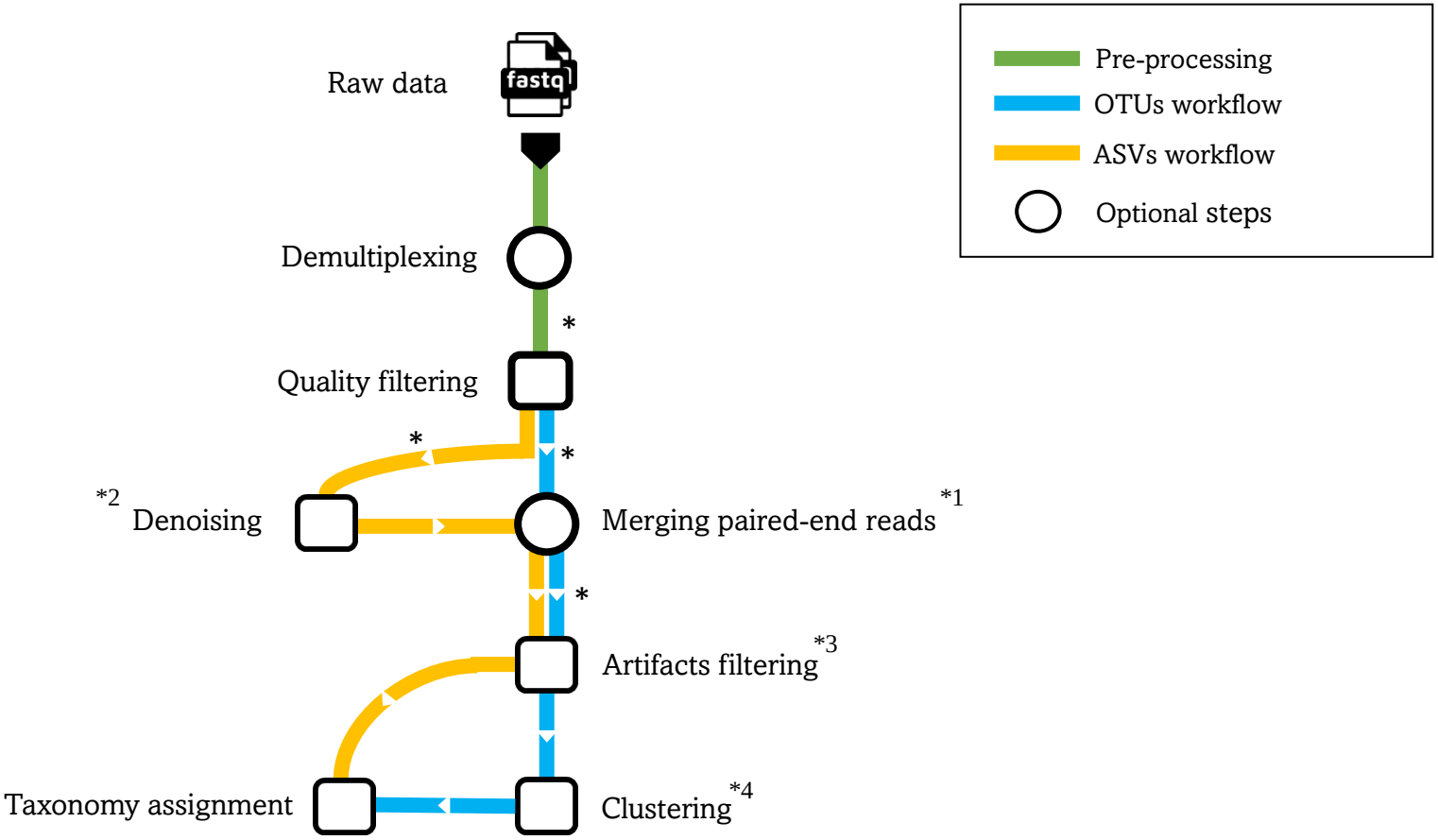
Zinger, L., Lionnet, C., Benoiston, A. S., Donald, J., Mercier, C., & Boyer, F. (2021). metabaR: an R package for the evaluation and improvement of DNA metabarcoding data quality. *Methods in Ecology and Evolution*, *12*(4), 586-592. https://doi.org/10.1111/2041-210X.13552

Figures

Figure 1. Examples of basic bioinformatics workflows for metabarcoding data. The workflow begins with demultiplexing, assigning reads to respective samples based on unique molecular identifiers. Next, quality filtering removes low-quality reads to reduce errors and improve reliability. Denoising algorithms identify and correct sequencing errors while preserving biological variation. For paired-end reads, merging combines forward and reverse reads into single sequences. Artifacts filtering removes biases introduced by sequencing artifacts like chimeras and NUMTs. Clustering groups sequences into OTUs or ASVs based on similarity,

reducing data complexity. Finally, taxonomic assignment is performed using reference databases and algorithms, enabling accurate identification of studied communities.
* Primer trimming between any of these steps can be applied.
*1 Only for paired-end data. May be performed before or after quality filtering.
*2 Error correction; formation of ASVs.
*3 Including chimera filtering, off-target gene removal (pseudogene removal, ITS extraction).
*4 Formation of OTUs/swarm-clusters.

Figure 2. Software for metabarcoding data bioinformatics processing categorized by input read type (paired-end, single-end (the tools in electric blue are capable of handling both paired-end and single-end reads)), software type (suite, pre-compiled pipeline), interface (CLI, GUI, Web, Galaxy web platform), produced feature type (OTU, ASV, swarm-cluster), and operating system (Linux, macOS, Windows).

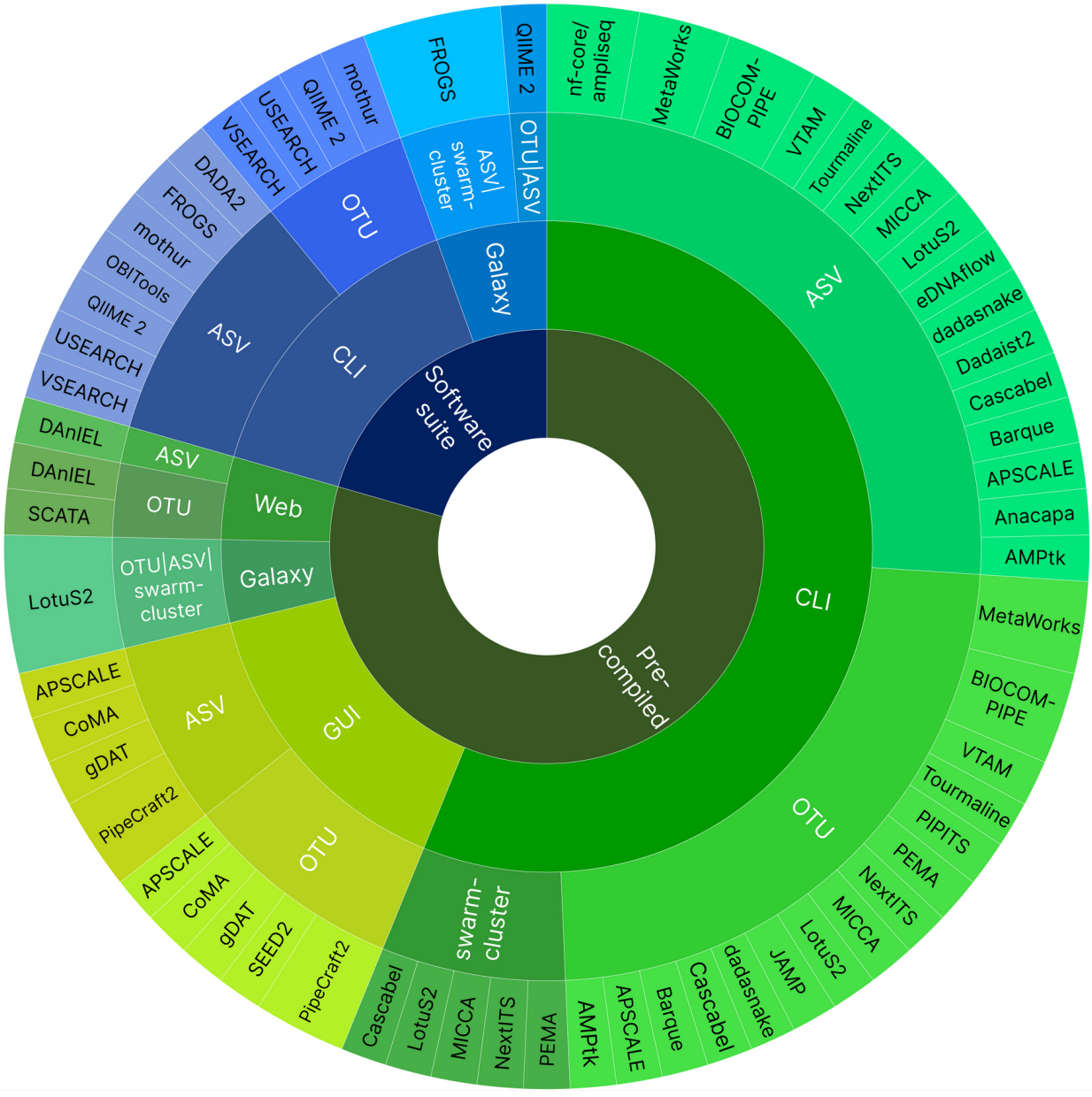


Single-end data

AMPtk Anacapa BIOCOM-PIPE Cascabel CoMA DADA2 dadasanke
eDNAflow FROGS gDAT JAMP LotuS2 MetaWorks MICCA mothur
NextITS nf-core/ampliseq OBITools PipeCraft2 QIIME 2 SCATA
SEED2 USEARCH VSEARCH Tourmaline VTAM

Paired-end data

APSCALE
Barque
Dadaist2
DAnIEL
PIPITS
PEMA



Linux

macOS

eDNAflow
dadasnake
MetaWorks
NextITS

AMPtk Barque BIOCOM-PIPE
Cascabel Dadaist2 JAMP
OBITools PIPITS VSEARCH

Anacapa APSCALE CoMA
DADA2 gDAT MICCA
mothur nf-core/ampliseq
PEMA PipeCraft2 USEARCH
Tourmaline VTAM

DAnIEL FROGS LotuS2 QIIME 2

SCATA

SEED2

Windows

Web-based (including Galaxy)