



TECHNICAL NOTE

PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes

Haris Zafeiropoulos ^{1,2,*}, Ha Quoc Viet¹, Katerina Vasileiadou^{1,3}, Antonis Potirakis¹, Christos Arvanitidis^{1,4}, Pantelis Topalis⁵, Christina Pavloudi¹ and Evangelos Pafilis¹

¹Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Former U.S. Base of Gournes P.O. Box 2214, 71003, Heraklion, Crete, Greece; ²Department of Biology, University of Crete, Voutes University Campus, Heraklion, Greece; ³Charles University, Department of Ecology, Faculty of Science, Viničná 7, CZ-12844, Prague, Czech Republic; ⁴LifeWatch ERIC, Plaza España SN, SECTOR II-III 41013, Seville, Spain and ⁵Institute of Molecular Biology and Biotechnology (IMBB), Foundation for Research and Technology (FORTH), Foundation for Research and Technology – Hellas, N. Plastira 100, GR-70013, Heraklion, Crete, Greece

*Correspondence address. Haris Zafeiropoulos, Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Hellenic Centre for Marine Research (HCMR), Heraklion, Greece. E-mail: haris-zaf@hcmr.gr  <http://orcid.org/0000-0002-4405-6802>

Abstract

Background: Environmental DNA and metabarcoding allow the identification of a mixture of species and launch a new era in bio- and eco-assessment. Many steps are required to obtain taxonomically assigned matrices from raw data. For most of these, a plethora of tools are available; each tool's execution parameters need to be tailored to reflect each experiment's idiosyncrasy. Adding to this complexity, the computation capacity of high-performance computing systems is frequently required for such analyses. To address the difficulties, bioinformatic pipelines need to combine state-of-the-art technologies and algorithms with an easy to get-set-use framework, allowing researchers to tune each study. Software containerization technologies ease the sharing and running of software packages across operating systems; thus, they strongly facilitate pipeline development and usage. Likewise programming languages specialized for big data pipelines incorporate features like roll-back checkpoints and on-demand partial pipeline execution. **Findings:** PEMA is a containerized assembly of key metabarcoding analysis tools that requires low effort in setting up, running, and customizing to researchers' needs. Based on third-party tools, PEMA performs read pre-processing, (molecular) operational taxonomic unit clustering, amplicon sequence variant inference, and taxonomy assignment for 16S and 18S ribosomal RNA, as well as ITS and COI marker gene data. Owing to its simplified parameterization and checkpoint support, PEMA allows users to explore alternative algorithms for specific steps of the pipeline without the need of a complete re-execution. PEMA was evaluated against both mock communities and previously published datasets and achieved results of comparable quality. **Conclusions:** A high-performance computing-based approach was used to develop PEMA; however, it can be used in personal computers as

Received: 18 November 2019; Revised: 5 January 2020; Accepted: 14 February 2020

© The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

well. PEMA's time-efficient performance and good results will allow it to be used for accurate environmental DNA metabarcoding analysis, thus enhancing the applicability of next-generation biodiversity assessment studies.

Keywords: pipeline; container; Docker; singularity; high performance computing; HPC; eDNA; metabarcoding

Background

Environmental DNA (eDNA) metabarcoding inaugurates a new era in bio- and eco-monitoring [1]. eDNA refers to genetic material obtained directly from environmental samples (soil, sediment, water, etc.) without any obvious signs of biological source material [2]. Metabarcoding is the combination of DNA taxonomy, based on taxa-specific marker genes (e.g., 16S ribosomal RNA [rRNA] for Bacteria and Archaea, cytochrome oxidase subunit 1 [COI] and 18S rRNA for Metazoa, ITS for Fungi), and high-throughput DNA sequencing technologies; thus, simultaneous identification of a mixture of organisms is attainable [3]. eDNA metabarcoding attempts to turn the page on the way biodiversity is perceived and monitored [3]. This combination is considered to be a potential holistic approach that, once standardized, allows for higher detection capacity and at a lower cost compared to conventional methods of biodiversity assessment. However, from the raw read sequence files to an amplicon study's results, the bioinformatics analysis required can be troublesome for many researchers.

Well-established pipelines are available to process metabarcoding data for the case of 16S and 18S rRNA marker genes and bacterial communities (e.g., mothur [4], QIIME 2 [5], LotuS [6]). However, certain limitations accompany each of these and occasionally they can be far from easy-to-use software. Moreover, there is a great need for similarly straightforward and benchmarked approaches for the analysis of other marker genes. With respect to the COI and ITS marker genes, a number of pipelines have been implemented, e.g., Barque [7], ScreenForBio [8], and PIPITS [9]. However, there is still need for a fast, flexible, easy-to-install, and easy-to-use pipeline for both COI and ITS marker genes.

The pipelines mentioned above, although entrenched, are still hindered by a series of hurdles. Among the most prominent are technical difficulties in installation and use, strict limitations in setting parameters for the algorithms invoked, and incompetence in partial re-execution of an analysis.

Moreover, given the computational demands of such analyses, access to high-performance computing (HPC) systems might be mandatory, e.g., to process studies with a large number of samples. This is timely given the ongoing investment of national and international efforts (e.g., [10]) to serve the broad biological community via commonly accessible infrastructures.

PEMA (Pipeline for Environmental DNA Metabarcoding Analysis) is an open source pipeline that bundles state-of-the-art bioinformatic tools for all necessary steps of amplicon analysis and aims to address the aforementioned issues. It is designed for paired-end sequencing studies and is implemented in the BDS [11] programming language. BDS's ad hoc task parallelism and task synchronization supports heavyweight computation, which PEMA inherits. In addition, BDS supports "checkpoint" files that can be used for partial re-execution and crash recovery of the pipeline. PEMA builds on this feature to serve tool and parameter exploratory customization for optimal metabarcoding analysis fine tuning. Switching effortlessly between (molecular) operational taxonomic unit ([M]OTU) clustering and amplicon sequence variant (ASV) inference algorithms is a pertinent example. Finally, via software containerization technologies such

as Docker [12] and Singularity [13], with the latter being HPC-centered, PEMA is distributed in an easy to download and install fashion on a range of systems, from regular computers to cloud or HPC environments.

From the biological perspective, monitoring biodiversity at all its different levels is of great importance. Because there is not a single marker gene to detect all taxa, researchers need to use different genes targeting each great taxonomy group separately [14]. To that end, PEMA supports the metabarcoding analysis of both prokaryotic communities, based on the 16S rRNA marker gene, and eukaryotic ones, based on the ITS (for Fungi) and COI and 18S rRNA (for Metazoa) marker genes [14].

As high-throughput sequencing (HTS) data become more and more accurate, ASVs, i.e., marker gene amplified sequence reads that differ in ≥ 1 nucleotide from each other, become easier to resolve [15]. The use of ASVs instead of OTUs has been suggested [15]; however, the choice of which approach to use should be based on each study's objective(s) [16].

PEMA supports both OTU clustering and ASV inference for all marker genes (see "OTU clustering vs ASV inference" in the "Results and Discussion" section). Two clustering algorithms, VSEARCH [17] and CROP [18], are used for the clustering of reads in (M)OTUs—the former for the case of the 16S/18S rRNA marker genes, the latter for the case of COI and ITS. Swarm v2 [19] allows ASV inference in all cases.

Taxonomic assignment is performed in an alignment-based approach, making use of the CREST LCAclassifier [20] and the Silva database [21] for the case of 16S and 18S rRNA marker genes; the Unite database [22] is used for the ITS gene. In the 16S marker gene case, phylogeny-based assignment is also supported, based on RAXML-ng [23], EPA-ng [24], and Silva [21]. For the COI marker gene, the RDPClassifier [25] and the MIDORI database [26] are used for the taxonomic assignment. In addition, ecological and phylogenetic analysis are facilitated via the "phyloseq" R package [27].

All the pipeline- and third-party module-controlling parameters are defined in a plain "parameter-value pair" text file. Its straightforward format eases the analysis fine tuning, complementary to the aforementioned checkpoint mechanism. A tutorial about PEMA and installation guidance can be found on PEMA's GitHub repository [28].

Implementation

PEMA's architecture comprises 4 main parts taking place in tandem (Fig. 1). A detailed description of the tools invoked by PEMA and their licenses is included in Additional File 1: Supplementary Methods.

Part 1: Quality control and pre-processing of raw data

First, FastQC [29] is used to obtain an overall read-quality summary; visual inspection of each sample's quality may recommend removing those insufficient quality, as well as samples with a low number of reads, and rerunning the analysis. To correct errors produced by the sequencer, PEMA incorporates a number of tools. Trimmomatic [30] implements a series of trim-

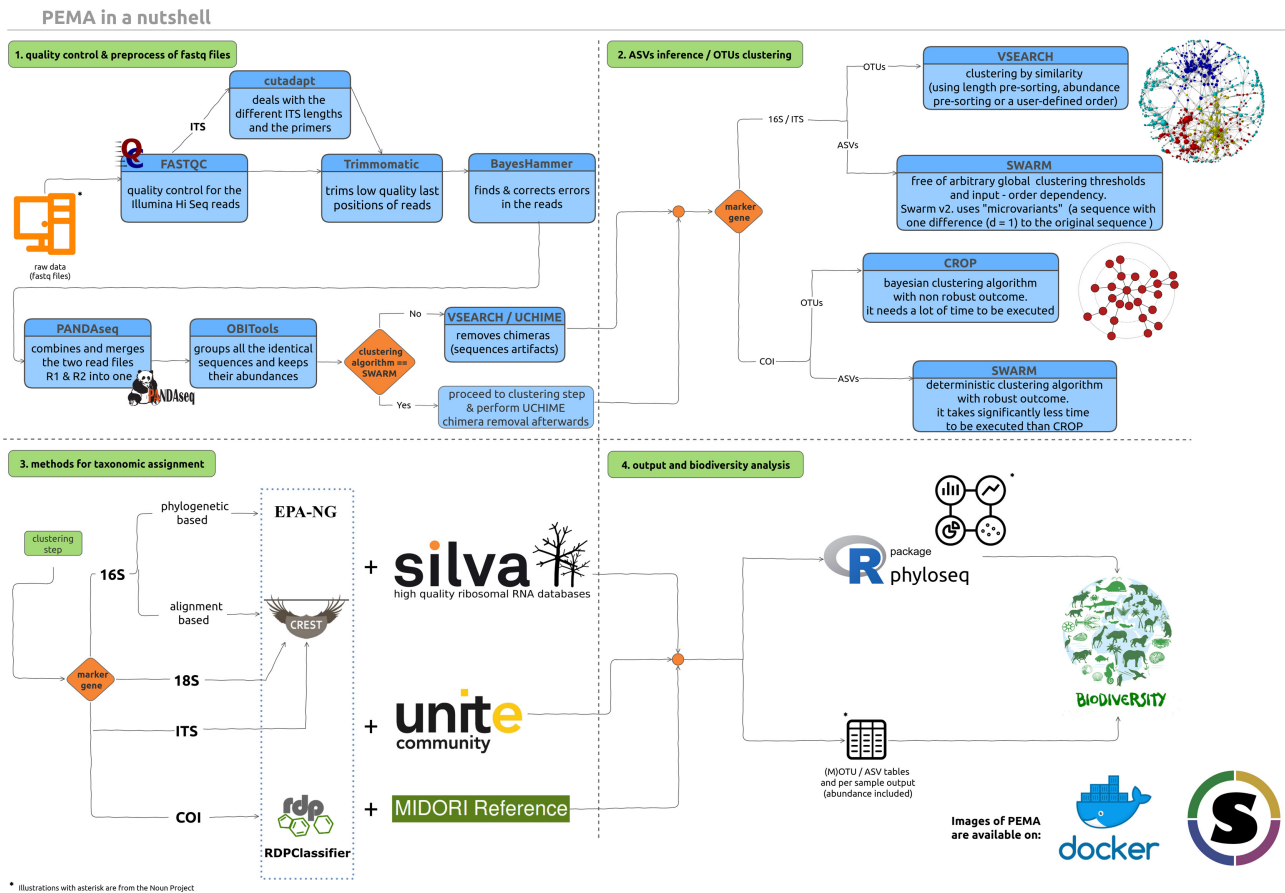


Figure 1: PEMA comprises 4 parts. The first step (top left) is the quality control and pre-processing of the Illumina sequencing reads. This step is common for both 16S rRNA and COI marker genes. The second step (top right) is the clustering of reads to (M)OTUs or their inferring to ASVs. The third step (bottom left) is the taxonomy assignment to the generated (M)OTUs/ASVs. In the fourth step (bottom right), the results of the metabarcoding analysis are provided to the user and visualized. *noun project icons by: ProSymbols (US), IconMark (PH), Nithinan Tatah (TH). clustering figure adapted from DOI: 10.7717/peerj.1420/fig-1.

ming steps, which either remove parts of the sequences corresponding to the adapters or the primers, trim and crop parts of the reads, or even remove a read completely, when it fails to reach the quality-filtering standards set by the user. Cutadapt [31] is used additionally for the case of ITS to address the variability in length of this marker gene (see Additional File 1: Supplementary Methods). BayesHammer [32], an algorithm of the SPAdes assembly toolkit [33], revises incorrectly called bases. PANDAseq [34] assembles the overlapping paired-end reads, and then the “obiuniq” program of OBITools [35] groups all the identical sequences in every sample, keeping track of their abundances. The VSEARCH package [17] is then invoked for chimera removal; however, if the Swarm v2 algorithm is selected, this step will be performed after the ASV inference (see next section).

Part 2: (M)OTU clustering and ASV inference

Quality-controlled and processed sequences are subsequently clustered into (M)OTUs or treated as input for inferring ASVs. For the case of 16S and 18S rRNA marker genes, VSEARCH [17] is used for OTU clustering, while ASVs can be identified by the Swarm v2 algorithm [19]. VSEARCH is an accurate and fast tool that can handle large datasets; at the same time it is a great alternative for USEARCH [36] because it is distributed under an open source license.

For the ITS and COI marker genes, CROP [18], an unsupervised probabilistic Bayesian clustering algorithm that models the clustering process using birth-death Markov chain Monte Carlo (MCMC), is used. The CROP clustering algorithm is adjusted by a series of parameters that need to be tuned by the user (namely, b , e , and z). These parameters depend on specific dataset properties such as the length and the number of reads. PEMA automatically adjusts b , e , and z by collecting such information and applying the CROP recommended parameter-setting rules [18]. ASV inference is conducted by Swarm v2 [19] in this case too.

Because the Swarm v2 algorithm is not affected by chimeras (F. Mahé, personal communication), when Swarm v2 is selected, chimera removal occurs after the clustering (see Additional File 1: Supplementary Methods: Swarm v2). This leads to a computational time gain as chimeras are sought among ASVs, instead of ungrouped reads.

Last, any singletons, i.e., sequences with only 1 read, occurring after the (M)OTU clustering or the ASV inference may be removed according to the user’s parameter settings.

Part 3: Taxonomy assignment

Alignment-based taxonomy assignment is supported for all marker gene analyses. In the case of the 16S/18S rRNA and ITS marker genes, the LCAClassifier algorithm of the CREST set of resources and tools [20] is used together with the Silva

[21] and the Unite [22] database, respectively, to assign taxonomy to the OTUs. Two versions of Silva are included in PEMA: 128 (29 September 2016) and 132 (13 December 2017). Because classifiers need first to be trained for each database they use, for future Silva [21] versions new PEMA versions will be available.

For the COI marker gene, PEMA uses the RDPClassifier [25] and the MIDORI reference database [26] to assign taxonomy of the MOTUs. The MIDORI database contains quality-controlled metazoan mitochondrial gene sequences from GenBank [37].

Intended primarily for studies from less explored environments, phylogeny-based assignment is available for 16S rRNA marker gene data. PEMA maps OTUs to a custom reference tree of 1,000 Silva-derived consensus sequences (created using RAXML-ng [23] and gappa [phat algorithm] [38], Fig. 2A). PaPaRa [39] and EPA-ng [24] combine the OTU clustering output and the reference tree to produce a phylogeny-aware alignment and map the 16S rRNA OTUs to the custom reference tree. Beyond the context of PEMA, users may visualize the output with tree viewers such as iTOL [40] (Fig. 2B).

Part 4: Ecological downstream analysis of the taxonomically assigned (M)OTU/ASV tables

PEMA's major output is either an (M)OTU or an ASV table with the assigned taxonomies and the abundances of each taxon in every sample. For each sample of the analysis, a subfolder containing statistics about the quality of its reads, as well as the taxonomies and their abundances, is also returned.

Via the “phyloseq” R package [27], downstream ecological analysis of the taxonomically assigned OTUs or ASVs is supported. This includes α - and β -diversity analysis, taxonomic composition, statistical comparisons, and calculation of correlations between samples.

When selected, in addition to the phyloseq [27] output, a multiple sequence alignment (MSA) and a phylogenetic tree of the OTU/ASVs retrieved can be returned; for the MSA, the MAFFT [41] aligner is invoked while the latter is built by RAXML-ng [23].

PEMA container-based installation

An easy way of installing PEMA is via its containers. A Dockerized PEMA version is available [42]. Singularity users can “pull” the PEMA image from [43]. Between the 2 containers, the Singularity-based one is recommended for HPC environments owing to Singularity's improved security and file accessing properties [44]. PEMA can also be found in the bio.tools (id: PEMA) and SciCrunch (PEMA, RRID:SCR.017676) databases. For detailed documentation, visit [28].

PEMA output

All PEMA-related files (i.e., intermediate files, final output, checkpoint files, and per-analysis parameters) are grouped in distinct (self-explanatory) subfolders per major PEMA pipeline step. In the last subfolder, i.e., subfolder 8, the results are further split into folders per sample. This eases further analysis both within the PEMA framework (e.g., partial re-execution for parameter exploration) and beyond. An extra subfolder is created when an ecological analysis via the “phyloseq” package has been selected.

Results and Discussion

Evaluation

To evaluate PEMA, 2 approaches were followed. First, PEMA was benchmarked against mock community datasets. Second, PEMA was used to analyse previously published datasets. PEMA's output was then compared with the original study outcome, as well as with the output of QIIME2, LotuS, Mothur, and Barque (where applicable).

Four mock communities, 1 for each marker gene, were used. With respect to the 16S rRNA marker gene, a mock community of Gohl et al. [45] with 20 different bacterial species was studied. Correspondingly, in the case of the 18S rRNA marker gene, a mock community of Bradley et al. [46] with 12 algal species was used; for the ITS, one of Bakker [47] including 19 different fungal taxa; and for the case of the COI marker gene, a mock community of Bista et al. [48] containing 14 metazoan species. More information on the mock communities, their original studies, and the results of PEMA for various combinations of parameters can be found in Additional File 2: Mock Communities.

Complementary to the mock community evaluation, 2 publicly available datasets from published studies were investigated through PEMA. For the 16S rRNA marker gene, the dataset reported by Pavlouidi et al. [49] was used; the original study aimed at investigating the sediment prokaryotic diversity along a transect river–lagoon–open sea. For the COI case, the dataset of Bista et al. [50] was used; this study investigated whether eDNA can be used for the accurate detection of chironomids (a taxonomic group of macroinvertebrates) in a freshwater habitat.

In both approaches, the respective .fastq files were downloaded from the European Nucleotide Archive (ENA) of the European Bioinformatics Institute ENA-(EBI) using “ENA File Downloader version 1.2” [51] and PEMA was run on the in-house HPC cluster.

All analyses were conducted on identical Dell M630 nodes (128 GB RAM, 20 physical Intel Xeon 2.60 GHz cores).

Mock community evaluation

PEMA was tested against mock communities. An evaluation of its accuracy must capture (i) how many of PEMA's predictions are true (i.e., the percent of correctly assigned taxa among all predicted taxa) and (ii) how many of the taxa existing in the mock community were recovered successfully by PEMA. The precision statistical metric was used to assess the former, and recall, the latter. In addition, the F1-score was used as a combined metric of both precision and recall. Precision is calculated as the ratio of true-positive results (TP) over the total number of true- (TP) and false-positive results (FP) predicted by a model, as follows: $\text{precision} = \text{TP}/(\text{TP} + \text{FP})$; recall is the ratio of TP over the total number of TP and false-negative results (FN): $\text{recall} = \text{TP}/(\text{TP} + \text{FN})$. The F1-score is the precision and recall harmonic mean and is calculated by means of the following formula: $\text{F1} = 2 \times (\text{precision} \times \text{recall})/(\text{precision} + \text{recall})$ [52].

Adequate accuracy was achieved when PEMA was used to recover the marker gene-specific mock communities at the genus level. Precision and recall scores of ~80% or more were observed, with 2 exceptions in precision but also 3 very high scores in recall. Overall the F1-scores ranged from 74% to 86%. A detailed description of the benchmark methodology and statistics analysis is given in Additional File 2: Mock Communities.

Detailed presentation of per-marker-gene-specific mock community recovery via PEMA is provided in the following sections. Several different sets of parameters were chosen for

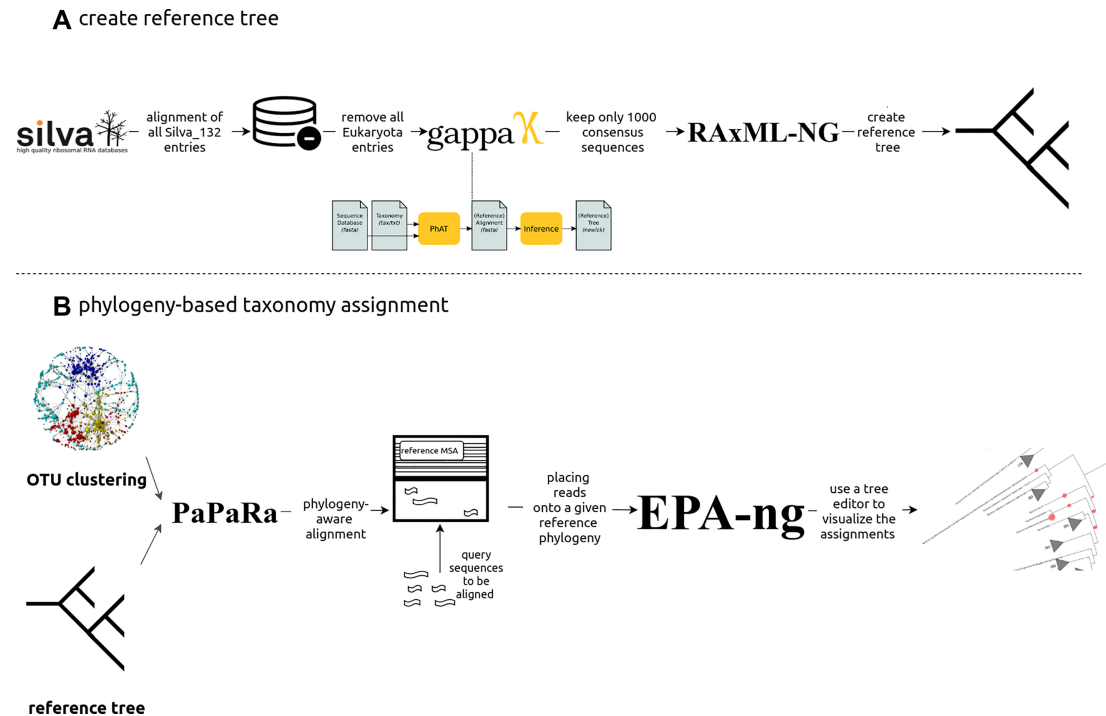


Figure 2: Phylogeny-based taxonomy assignment. A: Building a reference tree for the phylogeny-based taxonomy assignment to 16S rRNA marker gene OTUs: from the latest edition of Silva SSU, all entries referring to Bacteria and Archaea were used and using the “art” algorithm, 10,000 consensus taxa were kept. B: Using PaPaRa and the OTUs that come up from every analysis, an MSA was made and EPA-ng took over the phylogeny-based taxonomy assignment. *noun project icons by: Rockicon and A Beale.

each marker gene. Each marker gene has special features (e.g., length variability, sequence variability), and each Illumina run has its own intrinsic biases (e.g., primers used, PCR protocol); thus, parameter tuning plays a crucial part in metabarcoding analyses.

In an attempt to thoroughly analyse the sequence data from the mock communities, various sets of parameters were tested on the basis of the experimental details of the published studies but also in an exploratory way. Many different parameter settings were tested, especially for the steps of quality trimming of the reads and the OTU clustering/ASV inference. The differences in their output indicate how sensitive this method is, as well as the great need of a mock community in every metabarcoding study—both as a control but also as a “tuning system” for the parameter setting of the pipeline used.

16S rRNA

When PEMA was performed with the Swarm v2 algorithm ($d = 3$, strictness = 0.6) without removal of singletons, 18 of the 20 taxa were identified to the genus level and 3 of these even to the species level. There were 2 species that were not found in any of the PEMA runs. According to Gohl et al. [45], there was a discrepancy in the identification of those 2 species that was dependent on the amplification protocol used. It is worth mentioning that as d increases, taxa cannot be identified to species level at all; however, FP assignments decrease. Thus, when $d = 30$ and strictness = 0.6 for the KAPA samples, *Enterococcus* was not identified at all; however, PEMA finds its greatest F1 value (at the genus level, see Table 1) as the FP assignments returned are minimized. When PEMA was run using the VSEARCH clustering algorithm, high precision values were returned in all cases

Table 1: Summary benchmark of PEMA marker-gene-specific mock community recovery (precision)

Marker gene	Precision	Recall	F1
16S rRNA	0.81	0.85	0.83
18S rRNA	0.75	0.90	0.82
ITS	0.79	0.94	0.86
COI	0.62	0.93	0.74

(>0.79). However, the recall values were decreased when using Swarm v2 (0.65–0.68).

18S rRNA

When PEMA was performed using the Swarm v2 algorithm ($d = 1$, strictness = 0.5), 3 of 12 community members were identified to species level (*Isochrysis galbana*, *Nannochloropsis oculata*, and *Thalassiosira pseudonana*), 6 to genus, and the remaining 3 to class; the latter were all the green algae species (Chlorophyta) of the mock community. However, a better F1-score (0.82) was achieved when the class of Chlorophyceae was not found at all ($d = 1$, strictness = 0.3) because the FPs were decreased to only 1. When the VSEARCH algorithm was used, *I. galbana* was identified only to the genus level, the *Nannochloropsis* to the order level (Eustigmatales), and the *Poterioochromonas* genus to its class (Chrysophyceae).

ITS

When PEMA was performed using the Swarm v2 algorithm ($d = 20$) and targeting the ITS2 region, ASVs from 5 of the 19 species of the mock community were assigned to species level, 10 to

genus, 2 to family, and 2 to class level. Contrary to the study by Bakker [47], PEMA identified the genus *Chytriumyces* in all 3 samples, as well as the Ustilaginaceae family. Only 1 FP assignment was recorded. When the CROP algorithm was used, PEMA's output was less accurate; the *Fusarium* species contained in the mock community were not identified further than their family (Nectriaceae). As mentioned by Bakker [47], many reads deriving from the *Fusarium* spp. were not assigned to species level because of the quality-trimming step. In addition, a manually assembled reference database for the taxonomy assignment was used in the initial study, containing only sequences of the mock community species, which biased this step, making the results not directly comparable to our case.

COI

When PEMA was performed on the Bista et al. dataset [48] and using Swarm v2 ($d = 10$), it identified 12 of the 14 species included in the mock community. The sole non-identified species were *Bithynia leachii* and *Anisus vortex*. For *B. leachii* no entry exists in the MIDORI database, version MIDORI.LONGEST.1.1. However, the existence of another species of the genus *Bithynia* was recorded. With respect to *A. vortex*, PEMA returned a high abundance ASV assigned to the *Anisus* genus but with a low confidence level. PEMA managed to identify all the members of the mock community. This includes *Physa fontinalis*, which was originally not designed to be a member of the mock community but, as Bista et al. [48] explain, was recorded owing to cross-contamination. In the case of the COI marker gene, unique sequences with low abundances (singletons or doubletons) often lead to spurious MOTUs/ASVs. Thus, as shown in Additional File 2: Mock Communities, the FP assignments are decreased when these low-abundant sequences are removed; also, the abundance of the assignments (i.e., read counts) retrieved can indicate FP assignments. Thus, TP assignments occur in greater abundance, with hundreds or even thousands of reads—contrary to most of the FP results, whose abundance is <10 read counts. That is mostly for the case of the COI marker gene because eukaryotes are under study; eukaryotes have a great number of copies of this marker gene—different numbers of copies among the different species—and not just a single one as is almost always the case in bacteria. Therefore, assignments with such low abundances should be doubted as TP results in analyses on real datasets.

Comparison with existing software

PEMA's features were compared with those of mothur [4], QIIME 2 [5], LotuS [6], and Barque [7]. Table 2 presents a detailed comparison among the 4 tools' features in terms of marker gene support, diversity and phylogeny analysis capability, parameter setting and mode of execution, operation system availability, and HPC suitability. As shown, PEMA is equally feature-rich, if not richer in certain feature categories, compared with the other software packages. In particular, PEMA's support for COI marker gene studies is distinctive; 2 methods for taxonomy assignment are supported, and PEMA's easy parameter setting, step-by-step execution, and container distribution render it user and analysis friendly.

Evaluation on real datasets and against other tools

In the following sections, a comparative study on real datasets of the 16S rRNA and COI marker genes is presented. Analyses

using PEMA and the pipelines mentioned above that support each of these 2 marker genes were performed, both with multiple sets of parameters. It is typical for pipelines to invoke a variety of established tools. In many cases, a number of tools are common among different pipelines. Therefore, it is important to stress that such comparisons should not be taken into account strictly; declaring that one pipeline is better than another is not trivial. Potentials and limitations of both the pipelines and the metabarcoding method, as well as the importance of the role of the pipeline user, are underlined in the following sections.

16S rRNA marker gene analysis evaluation

To evaluate PEMA's performance, a comparative analysis of the Pavloudi et al. [49] dataset with mothur [4], QIIME 2 [5], LotuS [6], and PEMA was conducted.

It is known that the choice of parameters affects the output of each analysis; therefore, it is expected that different user choices might distort the derived outputs. For this reason and for a direct comparison of the pipelines, we have included all the commands and parameters chosen in the framework of this study in Additional File 1: Supplementary Methods. The results of the processing of the sequences by PEMA are presented in Table S1. All analyses were conducted on identical Dell M630 nodes (128 GB RAM, 20 physical Intel Xeon 2.60 GHz cores). LotuS, mothur, and QIIME 2 operated in a single-thread (core) fashion. PEMA, given the BDS intrinsic parallelization [11], operated with up to the maximum number of node cores (in this case 20).

The execution time and the reported OTU number of each tool are presented in Table 3. LotuS and PEMA resulted in a final number of OTUs comparable to that of Pavloudi et al. [49]. Clearly, owing to PEMA's parallel execution support, the analysis time can be significantly reduced (~ 1.5 hours in this case). The execution time depends on the parameters chosen for each software (see Additional File 1: Supplementary Methods).

Owing to the non-full overlap of the sequence reads, mothur resulted in an inflated number of OTUs; thus, it was excluded from further analyses. The results of all the pipelines were analysed with the phyloseq script that is provided with PEMA. The taxonomic assignment of the PEMA-retrieved OTUs is shown in Fig. 3. The phyla that were found in the samples are similar to the ones that were found in the original study [49]. Although the lowest number of OTUs was found in the marine station (Kal) (Supplementary Table S3), which is not in accordance with Pavloudi et al. [49], the general trend of a decreasing number of OTUs with increasing salinity was observed as in the original study (Supplementary Fig. S1). Notably, this result was not observed with the other tested pipelines (Supplementary Table S3). Furthermore, each of the pipelines resulted in a different taxonomic profile (Supplementary Figs S2–S4), with an extreme case of missing the order of Betaproteobacteriales (Supplementary Figs S5–S7).

Moreover, when the PERMANOVA analysis was run for the results of PEMA, LotuS, and DADA2, it was clear that the microbial community composition was significantly different in each of the 3 sampled habitats (i.e., river, lagoon, open sea) (PERMANOVA: F.Model = 7.0718, $P < 0.001$; F.Model = 6.5901, $P < 0.001$; F.Model = 2.2484, $P < 0.05$, respectively), which is in accordance with Pavloudi et al. [49]. However, this was not the case with Deblur (PERMANOVA: $P > 0.05$). Overall, PEMA's output is in accordance with the original study [49], and seen through this

Table 2: Comparison of the basic features of the different pipelines

Feature	LotuS	QIIME 2	mothur	Barque	PEMA
16S rRNA	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
18S rRNA	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
ITS	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
COI	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
diversity indices	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
alignment-based taxonomy assignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
phylogenetic-based taxonomy assignment	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
parameters assigned in the command line	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
parameters assigned through a text file	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
step-by-step execution	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
all steps in one go possible	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
available for any Operating System (Linux, OSX, Windows)	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
traditional application installation	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
available as a virtual machine	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
available as a container	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
available for HPC as a container (Singularity container)	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>

Table 3: OTU predictions and execution time for the different pipelines

Parameter	LotuS	mothur	QIIME 2		PEMA	Pavloudi et al. [49]
			Deblur	DADA2		
No. of OTUs	9,849	142,669	517	1,023	6,028	7,050
Execution time (h)	~9	~67*	2.5	~5	~1.5	~26

*(~56 if the reference database is already built).

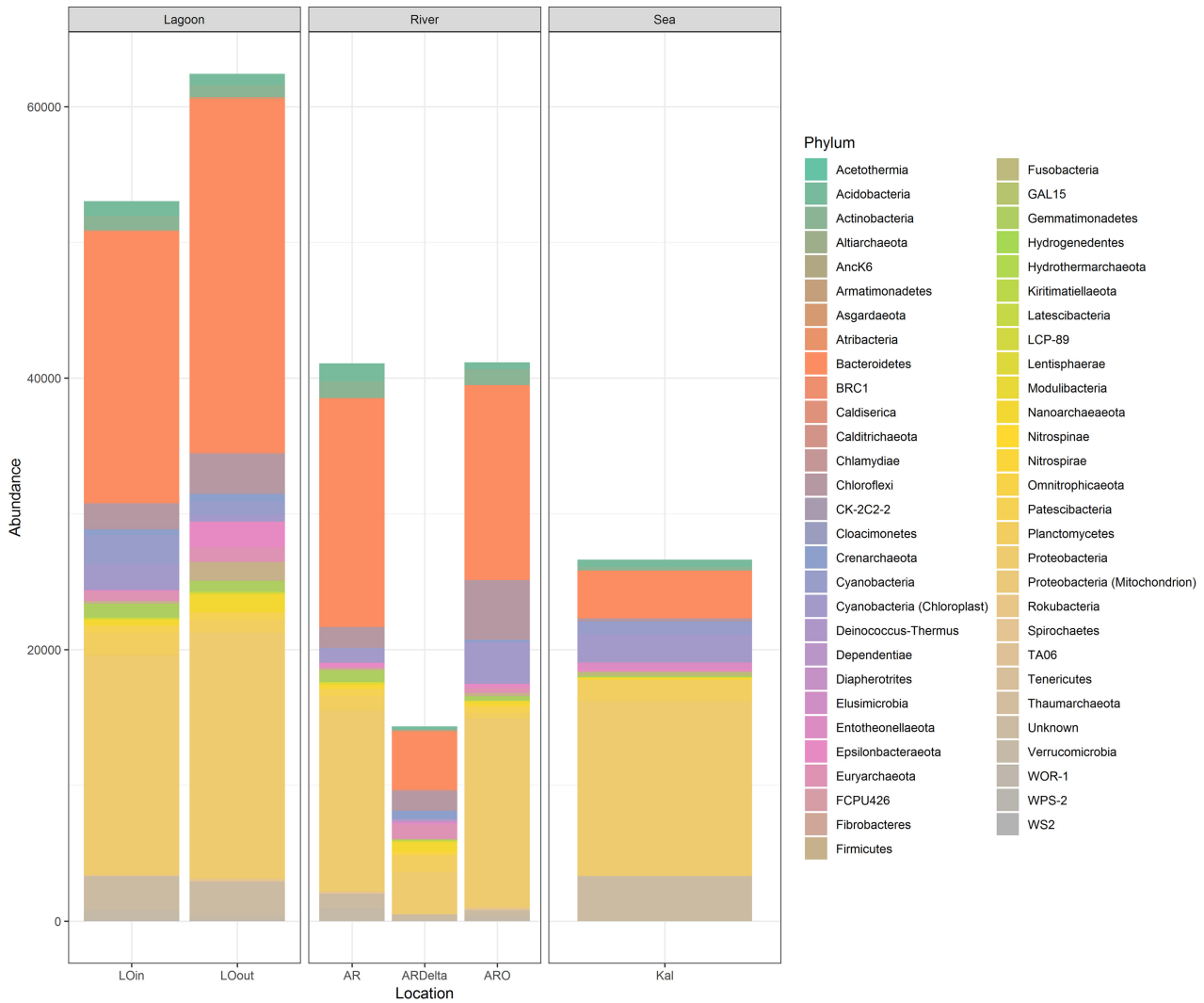


Figure 3: OTU bar plot at the phylum level. Bar plot depicting the taxonomy of the retrieved OTUs from PEMA for the dataset of Pavlouli et al. [49], at the phylum level for the case of the 16S marker gene. AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; LOin: Logarou station inside the lagoon; LOout: Logarou station in the channel connecting the lagoon to the gulf; Kal: Kalamitsi.

perspective PEMA performed equally well with the other tested pipelines, along with having the shortest execution time.

COI marker gene analysis evaluation

Bista et al. [50] created 2 COI libraries of different sizes: COIS (235-bp amplicon size) and COIF (658-bp amplicon size). The sequencing reads of COIS were selected for PEMA's evaluation; the COIF sequencing read pairs had no overlap so as to be merged and therefore were not considered appropriate for the analysis.

As previously, PEMA's performance was evaluated through a comparative analysis of the Bista et al. [50] dataset with Barque [7]; the commands and parameters chosen can be found in Additional File 1: Supplementary Methods. Regarding the creation of the MOTU table, in the Bista et al. [50] study VSEARCH [17] was used with a clustering at 97% similarity threshold. Afterwards, the BLAST+ (megablast) algorithm [53] was used against a manually created database including all NCBI GenBank COI sequences of length >100 bp (June 2015) while excluding environmental sequences and higher taxonomic level information [50].

As discussed in the publication, this approach resulted in 138 unique MOTUs of which 73 were assigned to species level. For PEMA's evaluation, the chosen clustering algorithm was Swarm v2, using different options for the cluster radius (d) parameter (Table 4); according to Mahé et al. [19], this is the most important parameter because it affects the number of MOTUs that are being created. The resulting MOTUs were classified against the MIDORI reference database [26] using RDPClassifier [25]. The results of the processing of the sequences are reported in Supplementary Table S3. For the case of Barque, the BOLD Database was used [54].

As shown in Table 4, PEMA resulted in 83 species-level MOTUs with a cluster radius (d) of 2, which is similar to the findings of the published study (i.e., 73 species). Although both the clustering algorithm and the taxonomy assignment methods were different between the original [50] and the present study, the results regarding the number of unique species present in the samples are in agreement to a considerable extent.

The computational time required by PEMA for the completion of the analysis is also reported in Table 4. Regardless of

Table 4: PEMA's^a output and execution time

Parameter	$d = 1$	$d = 2$	$d = 3$	$d = 10$	$d = 13$
MOTUs after pre-process and clustering steps	83,791	59,833	33,227	7,384	4,829
MOTUs after chimera removal	80,347	57,863	32,539	7,339	4,796
Non-singleton MOTUs	6,381	4,947	2,658	1,914	1,634
Assigned species	62	83	86	86	84
Execution time (h)	2:01:35	2:09:49	1:51:44	2:17:26	2:31:15

^aPEMA's output and execution time (using a 20-core node) for different values of Swarm's d parameter.

the value of the d parameter, all analyses were completed in ~2 hours, i.e., fast enough to allow parameter testing and customization. Regarding Barque, the analysis resulted in the identification of 51 species-level MOTUs and was concluded in 15 minutes. This difference is due to the error correction step of PEMA (BayesHammer algorithm [32]), which plays an important part in the enhanced results that PEMA returns, but it also requires a certain computational time; Barque does not have an analogous step, and therefore its overall execution time is shorter.

PEMA performed better than Barque at identifying taxa that were included in the positive control contents of the published study (Table 5).

OTU clustering vs ASV inference

There is an ongoing discussion about whether ASVs exceed OTUs. The strongest argument to this end is that ASVs are real biological sequences. Hence, they can be compared between different studies in a straightforward way; considered as consistent labels. In comparison, *de novo* OTUs are constructed, or “clustered,” with respect to the emergent features of each specific dataset. Therefore, OTUs defined in 2 different datasets cannot be directly compared.

However, the OTU concept is not compulsorily related to the clustering approach; it is widely used to describe results based on its biological meaning but it does not imply clustering. In addition, according to Callahan et al. [15], “ASV methods infer the biological sequences in the sample prior to the introduction of amplification and sequencing errors, and distinguish sequence variants differing by as little as one nucleotide.” As a result, ASVs could be considered as OTUs of higher resolution.

It is due to this concept confusion that algorithms whose rationale is considerably closer to the variant-based approach are still considered as OTU clustering algorithms [15]. Swarm v2 produces all possible “microvariants” of an amplicon to implement an exact-string comparison [19]. Furthermore, real biological sequences, “clouds of microvariants,” are produced as its output, which can be used for comparisons between different studies. Thus, Swarm v2 can be considered as an ASV-inferring algorithm.

Traditional clustering methods have certain limitations such as arbitrary global clustering thresholds and centroid selection because they depend on the input order and are time-consuming, etc. [55], which variant-based approaches manage to address. However certain algorithms for OTU clustering such as VSEARCH have been proven to be especially reliable, and they are widely used by many researchers. Furthermore, ASVs intend to improve taxonomic resolution; however, a vast number of inferred ASVs [56] can lead to inflation of diversity estimates, especially in the case of microbial communities, thus making the analysis even more complicated.

ASV or OTU approaches are supported by PEMA, although we have found that similar ecological results are produced by both these methods, as also suggested by Glassman and Martiny [57].

Beyond environmental ecology, ongoing and future work

PEMA is mainly intended to support eDNA metabarcoding analysis and be directly applicable to next-generation biodiversity/ecological assessment studies. Given that community composition analysis may also serve additional research fields, e.g., microbial pathology, the potential impact of such pipelines is expected to be much higher. Ongoing PEMA work focuses on serving a wide scientific audience and on making it applicable to more types of studies. The easy set-up and execution of PEMA allows users to work closely with national and European HPC/e-infrastructures (e.g., ELIXIR Greece [58], LifeWatch ERIC [59], EM-BRC ERIC [60]). To that end and in a mid-term perspective, a CWL version of PEMA will be explored. The aim of this effort is to reach out to a wider scientific audience and address both their ongoing as well as future analysis needs.

By supporting the analysis of the most commonly used marker genes for Bacteria and Archaea (16S rRNA), Fungi (ITS), and Metazoa (COI/18S rRNA), a holistic biodiversity assessment approach is now possible through PEMA and eDNA metabarcoding; although, from a mid-term perspective, it is our intention to allow ad hoc and in-house databases to be used as reference for the taxonomy assignment.

Conclusions

PEMA is an accurate, execution-friendly and fast pipeline for eDNA metabarcoding analysis. It provides a per-sample analysis output, different taxonomy assignment methods, and graphics-based biodiversity/ecological analysis. This way, in addition to (M)OTU/ASV calling, it provides users with both an informative study overview and detailed result snapshots.

Thanks to a nominal number of installation and execution commands required for PEMA to be set and run, it is considered essentially user friendly. In addition, PEMA's strategic choice of a single parameter file, implementation programming language, and multiple container-type distribution grant it speed (running in parallel), on-demand partial pipeline enactment, and provision for HPC-system-based sharing.

All the aforementioned features render PEMA attractive for biodiversity/ecological assessment analyses. By supporting the analysis of the most commonly used marker genes for Prokaryotes (Bacteria and Archaea), as well as Eukaryotes (Fungi and Metazoa), PEMA allows assessment of biodiversity in different levels of biodiversity. Applications may mainly concern environmental ecology, with possible extensions to such fields as microbial pathology and gut microbiome, in line with modern research needs, from low volume to big data.

Table 5: Comparison of the taxonomy of retrieved MOTUs among PEMA, Barque, and the positive controls of Bista et al. [50]

Barque	PEMA	Bista et al. [50]
<i>Ablabesmyia monilis</i> *	<i>Ablabesmyia monilis</i> <i>Crangonyx pseudogracilis</i> <i>Radix</i> sp.* Chironomidae sp.* <i>Ancyclus</i> sp.** <i>Athripsodes aterrimus</i> , <i>Athripsodes cinereus</i> **	<i>Ablabesmyia monilis</i> <i>Crangonyx pseudogracilis</i> <i>Radix</i> sp. Chironomidae sp. <i>Ancyclus fluviatilis</i> <i>Athripsodes albifrons</i>
<i>Chironomus anthracinus</i> **	<i>Chironomus</i> sp., <i>Chironomus anthracinus</i> , <i>Chironomus pseudothummi</i> , <i>Chironomus riparius</i> **	<i>Chironomus tentans</i>
<i>Polypedilum sordens</i> **		<i>Polypedilum nubeculosum</i>
<i>Athripsodes aterrimus</i> **		<i>Athripsodes albifrons</i>

*: Taxonomies identical to the published study (species level).

** : Taxonomies identical to the published study (genus level).

Availability of Supporting Source Code and Requirements

Project name: PEMA

Project home page: <https://github.com/hariszaf/pema>

Dockerized version: <https://hub.docker.com/r/hariszaf/pema>

Singularity image: <https://singularity-hub.org/collections/2295>

Operating system(s): Platform independent

Programming language: BigDataScript

Other requirements: Singularity (in case of HPC use)

License: GNU GPLv3. For third-party components separate licenses apply. See Additional File 1 for a list of tools invoked by PEMA and their respective licenses.

bio.tools id: PEMA

RRID:SCR_017676

Availability of Supporting Data and Materials

The sequence data that support the findings of this study, with respect to the mock community-based evaluation, are available in the European Nucleotide Archive (ENA) with the following study accession numbers—for the 16S, 18S rRNA, ITS, and COI marker genes, respectively:

PRJNA305443 (<https://www.ebi.ac.uk/ena/browser/view/PRJNA305443>),

PRJNA314977 (<https://www.ebi.ac.uk/ena/browser/view/PRJNA314977>),

PRJNA377530 (<https://www.ebi.ac.uk/ena/browser/view/PRJNA377530>), and

PRJEB23036 (<https://www.ebi.ac.uk/ena/browser/view/PRJEB23036>)

The real datasets used are also available in ENA:

PRJEB20211 (<http://www.ebi.ac.uk/ena/data/view/PRJEB20211>) and

PRJEB13009 (<https://www.ebi.ac.uk/ena/data/view/PRJEB13009>).

An archived version of the code and supporting data is also available via the GigaScience database GigaDB [61].

Additional Files

Additional File 1: Supplementary Methods: Description of tools invoked by PEMA and their licences. Description of the commands, along with their parameters, used to run PEMA, mothur, LotuS, and QIIME 2.

Additional File 2: Mock Communities: Details about the mock communities chosen and their corresponding studies, as well

as the returned output of PEMA for each for a number of sets of parameters.

Supplementary Table S1: Number of sequences after each pre-processing step for the case of 16S rRNA gene.

Supplementary Table S2: Diversity indices of the samples.

Supplementary Figure S1: Linear regression between the number of OTUs (averaged per sampling station) and the salinity of the sampling stations. L: Lagoon; S: Sea; R: River; AR: Arachthos; ARO: Arachthos Neochori; ARDelta: Arachthos Delta; LOin: Logarou station inside the lagoon; LOout: Logarou station in the channel connecting the lagoon to the gulf; Kal: Kalamitsi.

Supplementary Figure S2: Bar plot depicting the taxonomy of the retrieved OTUs from LotuS at the phylum level.

Supplementary Figure S3: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using Deblur at the phylum level.

Supplementary Figure S4: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using DADA2 at the phylum level.

Supplementary Figure S5: Bar plot depicting the taxonomy of the retrieved OTUs from LotuS at the class of Betaproteobacteriales.

Supplementary Figure S6: Bar plot depicting the taxonomy of the retrieved OTUs from QIIME 2 using Deblur at the class of Betaproteobacteriales.

Supplementary Figure S7: Bar plot depicting the taxonomy of the retrieved OTUs from PEMA at the class of Betaproteobacteriales.

Supplementary Table S3: Number of sequences after each pre-processing step for the case of COI, dataset from Bista et al. [50].

Abbreviations

BDS: BigDataScript; bp: base pairs; COI: cytochrome oxidase subunit 1; CREST: Classification Resources for Environmental Sequence Tags; CROP: Clustering 16S rRNA for OTU Prediction; CWL: Common Workflow Language; eDNA: environmental DNA; FN: false negative; FP: false positive; HCMR: Hellenic Centre for Marine Research; HPC: high-performance computing; iTOL: Interactive Tree of Life; MAFFT: Multiple Alignment using Fast Fourier Transform; MCMC: Markov chain Monte Carlo; MOTU: molecular operational taxonomic unit (used for eukaryotes); MSA: multiple sequence alignment; NCBI: National Center for Biotechnology Information; OTU: operational taxonomic unit (used for prokaryotes); PaPaRa: Parsimony-based

Phylogeny-Aware Read Alignment; PEMA: Pipeline for Environmental DNA Metabarcoding Analysis; PERMANOVA: permutational multivariate analysis of variance; RAM: random access memory; RAxML: Randomized Axelerated Maximum Likelihood; rRNA: ribosomal RNA; SPAdes: St. Petersburg genome assembler; SSU: small subunit; TP: true positive.

Competing Interests

The authors declare that they have no competing interests.

Funding

This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under grant agreement No. 241 (PREGO project) and from the project RECONNECT (MIS 5017160) financed by the Transnational Cooperation Programme Interreg V-B “Balkan-Mediterranean 2014-2020” and co-funded by the European Union and national funds of the participating countries. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

H.Z. conceived and designed the pipeline, performed its containerization, analysed and interpreted the data, wrote the paper, prepared figures and/or tables, and reviewed drafts of the paper. H.Q.V. offered support in the HPC preparation and set-up and in third-party component use. K.V. and C.A. conceived the idea and reviewed drafts of the paper. P.T. conceived the idea, proposed the use of the programming language, and reviewed drafts of the paper. C.P. conceived the idea, prepared figures and/or tables, and reviewed drafts of the paper. AP offered support in HPC and in third-party components. E.P. conceived the idea, assisted with programming and set-up, and reviewed drafts of the paper. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the Information technology (IT) group of HCMR and especially Mr. Stelios Ninidakis, Mr. Georgios Tsamis, and Mr. Dimitris Sidirokastritis for their help and support during cluster maintenance and installation of third-party software. They also thank Dr. Christos A. Christakis (ORCID iD: 0000-0002-7075-0996) for his valuable feedback on ecological analysis usefulness aspects.

This research was supported in part through computational resources provided by IMBBC (Institute of Marine Biology, Biotechnology and Aquaculture) of the HCMR (Hellenic Centre for Marine Research). Funding for establishing the IMBBC HPC has been received by the MARBIGEN (EU Regpot) project, Life-WatchGreece RI, and the CMBR (Centre for the Study and Sustainable Exploitation of Marine Biological Resources) RI.

References

- Pavan-Kumar A, Gireesh-Babu P, Lakra WS. DNA metabarcoding: a new approach for rapid biodiversity assessment. *J Cell Sci Mol Biol* 2015;2(1):111.
- Thomsen PF, Willerslev E. Environmental dna—an emerging tool in conservation for monitoring past and present biodiversity. *Biol Conserv* 2015;183:4–18.
- Ji Y, Ashton L, Pedley SM, et al. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecol Lett* 2013;16(10):1245–57.
- Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009;75:7537–41.
- Bolyen E, Rideout JR, Dillon MR, et al. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Preprints* 2018;6:e27295v2.
- Hildebrand F, Tadeo R, Voigt AY, et al. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* 2014;2:30.
- Normandeau E. Environmental DNA metabarcoding analysis. <https://github.com/enormandeau/barque>. Accessed 10 November 2019.
- Axtner J, Crampton-Platt A, Hoerig LA, et al. An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. *Gigascience* 2019;8(4):giz029.
- Gweon HS, Oliver A, Taylor J, et al. PIPITS: an automated pipeline for analyses of fungal internal transcribed spacer sequences from the Illumina sequencing platform. *Methods Ecol Evol* 2015;6(8):973–80.
- European Strategy Forum on Research Infrastructures Innovation Working Group. Innovation-oriented cooperation of Research Infrastructures. Vol. 3. ESFRI Scripta. 2018. ISBN Print: 978-88-943243-0-3.
- Cingolani P, Sladek R, Blanchette M. BigDataScript: a scripting language for data pipelines. *Bioinformatics* 2014;31:10–16.
- Rad BB, Bhatti HJ, Ahmadi M. An introduction to Docker and analysis of its performance. *Int J Comput Sci Netw Secur* 2017;17:228.
- Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. *PLoS One* 2017;12:e0177459.
- Coissac E, Riaz T, Puillandre N. Bioinformatic challenges for DNA metabarcoding of plants and animals. *Mol Ecol* 2012;21(8):1834–47.
- Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017;11(12):2639.
- Pauvert C, Buée M, Laval V, et al. Bioinformatics matters: the accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecol* 2019;41:23–33.
- Rognes T, Flouri T, Nichols B, et al. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 2016;4:e2584.
- Hao X, Jiang R, Chen T. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 2011;27:611–8.
- Mahé F, Rognes T, Quince C, et al. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 2015;3:e1420.
- Lanzén A, Jørgensen SL, Huson DH, et al. CREST—Classification Resources for Environmental Sequence Tags. *PLoS One* 2012;7:e49334.
- Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 2013;41:D590–6.

22. Nilsson RH, Larsson KH, Taylor AF, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 2018;**47**(D1):D259–64.
23. Kozlov AM, Darriba D, Flouri T, et al. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 2019;**35**(21):4453–5.
24. Barbera P, Kozlov AM, Czech L, et al. EPA-ng: massively parallel evolutionary placement of genetic sequences. *Syst Biol* 2018;**68**:365–9.
25. Wang Q, Garrity GM, Tiedje JM, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;**73**:5261–7.
26. Machida RJ, Leray M, Ho SL, et al. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Sci Data* 2017;**4**:170027.
27. McMurdie JP, Holmes S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 2013;**8**:e61217.
28. PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S rRNA, ITS and COI marker genes. <https://github.com/hariszaf/pema>. Accessed on, November 2019.
29. Andrews S. FastQC. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 8 July 2019.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.
31. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**(1):10–2.
32. Nikolenko SI, Korobeynikov AI, Alekseyev MA. Bayeshammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* 2013;**14**:S7.
33. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;**19**:455–77.
34. Masella AP, Bartram AK, Truszkowski JM, et al. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 2012;**13**:31.
35. Boyer F, Mercier C, Bonin A, et al. OBITools: a UNIX-inspired software package for DNA metabarcoding. *Mol Ecol Resour* 2016;**16**:176–82.
36. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;**26**(19):2460–1.
37. Benson DA, Cavanaugh M, Clark K, et al. GenBank. *Nucleic Acids Res* 2018;**46**:D41–47.
38. Czech L, Barbera P, Stamatakis A. Methods for automatic reference trees and multilevel phylogenetic placement. *Bioinformatics* 2018;**35**:1151–8.
39. Berger SA, Stamatakis A. PaPaRa 2.0: a vectorized algorithm for probabilistic phylogeny-aware alignment extension. Heidelberg Institute for Theoretical Studies 2012. <https://cme.its.org/exelixis/web/software/papara/index.html>.
40. Letunic I, Bork P. Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 2006;**23**:127–8.
41. Katoh K, Misawa K, Kuma KI, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 2002;**30**:3059–66.
42. PEMA: flexible Pipeline for eDNA Metabarcoding Analysis of the 16S/18S rRNA, ITS & COI marker genes. <https://hub.docker.com/r/hariszaf/pema>. Accessed on, November 2019.
43. <https://singularity-hub.org/collections/2295>. Accessed on, November 2019.
44. Chavez J. Singularity: a “Docker” for HPC environments. <https://dev.to/grokode/singularity--a-docker-for-hpc-environments-i6p>. Accessed on, 8 July 2019.
45. Gohl DM, Vangay P, Garbe J, et al. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat Biotechnol* 2016;**34**(9):942.
46. Bradley IM, Pinto AJ, Guest JS. Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Appl Environ Microbiol* 2016;**82**(19):5878–91.
47. Bakker MG. A fungal mock community control for amplicon sequencing experiments. *Mol Ecol Resour* 2018;**18**(3): 541–56.
48. Bista I, Carvalho GR, Tang M, et al. Performance of amplicon and shotgun sequencing for accurate biomass estimation in invertebrate community samples. *Mol Ecol Resour* 2018;**18**(5):1020–34.
49. Pavlouli C, Kristoffersen JB, Oulas A, et al. Sediment microbial taxonomic and functional diversity in a natural salinity gradient challenge Remane’s “species minimum” concept. *PeerJ* 2017;**5**:e3687.
50. Bista I, Carvalho GR, Walsh K, et al. Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. *Nat Commun* 2017;**8**: 14087.
51. Harrison PW, Alako B, Amid C, et al. The European Nucleotide Archive in 2018. *Nucleic Acids Res* 2018;**47**: D84–8.
52. Ting KM. Precision and recall. In: Sammut C, Webb GI, eds. *Encyclopedia of Machine Learning*. Boston, MA: Springer, 2011.
53. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
54. Ratnasingham S, Hebert PD. BOLD: the barcode of life data system (<http://www.barcodinglife.org>). *Mol Ecol Notes* 2007;**7**(3):355–64.
55. Mahé F, Rognes T, Quince C, et al. Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* 2014;**2**:e593.
56. Fierer N, Brewer T, Choudoir M. Lumping versus splitting – is it time for microbial ecologists to abandon OTUs? 2017. <http://fiererlab.org/2017/05/02/lumping-versus-splitting-is-it-time-for-microbial-ecologists-to-abandon-otus/>. Accessed on, 20 December 2019.
57. Glassman SI, Martiny JB. BROADSCALE ecological patterns are robust to use of exact sequence variants versus operational taxonomic units. *MSphere* 2018;**3**(4):e00148–18.
58. ELIXIR-GR. <https://www.elixir-greece.org/>. Accessed on, 8 July 2019.
59. LifeWatch-ERIC. <https://www.lifewatch.eu/>. Accessed on, 8 July 2019.
60. EMBRC. <http://www.embrc.eu/>. Accessed on, 8 July 2019.
61. Zafeiropoulos H, Quoc VH, Vasileiadou K, et al. Supporting data for “PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S rRNA, ITS, and COI marker genes.” GigaScience Database 2020. <http://dx.doi.org/10.5524/100715>. Accessed on, November 2019.