ARPHA Conference Abstracts 4: e64902
doi: 10.3897/aca.4.e64902

Conference Abstract

PEMA v2: addressing metabarcoding bioinformatics analysis challenges

Haris Zafeiropoulos[‡,§], Christina Pavloudi[§], Evangelos Pafilis[§]

‡ University of Crete, Heraklion, Greece
§ Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Heraklion, Crete, Greece

Corresponding author: Haris Zafeiropoulos (haris-zaf@hcmr.gr)

Received: 23 Feb 2021 | Published: 04 Mar 2021

Citation: Zafeiropoulos H, Pavloudi C, Pafilis E (2021) PEMA v2: addressing metabarcoding bioinformatics analysis challenges. ARPHA Conference Abstracts 4: e64902. https://doi.org/10.3897/aca.4.e64902

Abstract

Environmental DNA (eDNA) and metabarcoding have launched a new era in bio- and eco-assessment over the last years (Ruppert et al. 2019). The simultaneous identification, at the lowest taxonomic level possible, of a mixture of taxa from a great range of samples is now feasible; thus, the number of eDNA metabarcoding studies has increased radically (Deiner and 2017). While the experimental part of eDNA metabarcoding can be rather challenging depending on the special characteristics of the different studies, computational issues are considered to be its major bottlenecks. Among the latter, the bioinformatics analysis of metabarcoding data and especially the taxonomy assignment of the sequences are fundamental challenges.

Many steps are required to obtain taxonomically assigned matrices from raw data. For most of these, a plethora of tools are available. However, each tool's execution parameters need to be tailored to reflect each experiment's idiosyncrasy; thus, tuning bioinformatics analysis has proved itself fundamental (Kamenova 2020). The computation capacity of high-performance computing systems (HPC) is frequently required for such analyses. On top of that, the non perfect completeness and correctness of the reference taxonomy databases is another important issue (Loos et al. 2020).

Based on third-party tools, we have developed the Pipeline for Environmental Metabarcoding Analysis (PEMA), a HPC-centered, containerized assembly of key

metabarcoding analysis tools. PEMA combines state-of-the art technologies and algorithms with an easy to get-set-use framework, allowing researchers to tune thoroughly each study thanks to roll-back checkpoints and on-demand partial pipeline execution features (Zafeiropoulos 2020).

Once PEMA was released, there were two main pitfalls soon to be highlighted by users. PEMA supported 4 marker genes and was bounded by specific reference databases. In this new version of PEMA the analysis of any marker gene is now available since a new feature was added, allowing classifiers to train a user-provided reference database and use it for taxonomic assignment. Fig. 1 shows the taxonomy assignment related PEMA modules; all those out of the dashed box have been developed for this new PEMA release. As shown, the RDPClassifier has been trained with Midori reference 2 and has been added as an option, classifying not only metazoans but sequences from all taxonomic groups of Eukaryotes for the case of the COI marker gene. A PEMA documentation site is now also available. PEMA.v2 containers are available via the DockerHub and SingularityHub as well as through the Elixir Greece AAI Service. It has also been selected to be part of the LifeWatch ERIC Internal Joint Initiative for the analysis of ARMS data and soon will be available through the Tesseract VRE.

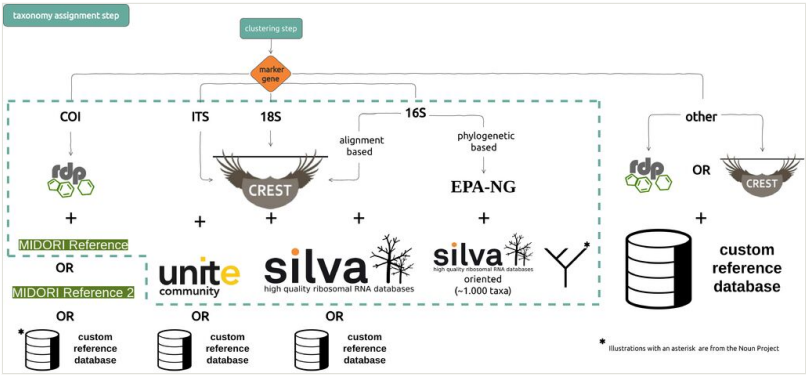

Figure 1. doi

Taxonomy assignment related features on initial version of PEMA and PEMA.v2. Custom databases can be now used for the taxonomy assignment of the four marker genes initially supported by PEMA. The analysis of further marker genes is now supported by providing PEMA with corresponding reference databases in the appropriate format to train either the CREST or the RDPClassifier.

Keywords

metabarcoding, pipeline, reference database, marker genes, 16S/18S rRNA, COI, ITS, containers, HPC

Presenting author

Haris Zafeiropoulos

Presented at

1st DNAQUA International Conference (March 9-11, 2021)

Funding program

This project has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRI), under grant agreement No. 241 (PREGO project), from the project RECONNECT (MIS 5017160) financed under 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. It has been also supported by the "ELIXIR-GR: Managing and Analysing Life Sciences Data (MIS: 5002780)" project co-financed by Greece and the European Union - European Regional Development Fund.

References

- Deiner K, (2017) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Molecular ecology 26 (21).
- Kamenova S (2020) A flexible pipeline combining clustering and correction tools for prokaryotic and eukaryotic metabarcoding. Peer Community in Ecology.
- Loos, M. L, Nijland R (2020) Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. Molecular Ecology.
- Ruppert K, Kline R, Rahman MS (2019) Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Global Ecology and Conservation 17.
- Zafeiropoulos H, et al. (2020) PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. GigaScience 9 (3).