Bioinformatics, 37(22), 2021, 4253–4254
doi: 10.1093/bioinformatics/btab431
Advance Access Publication Date: 12 June 2021
Applications Note

Sequence analysis

ITSoneWB: profiling global taxonomic diversity of eukaryotic communities on Galaxy

Marco A. Tangaro[1], Giuseppe Defazio [2], Bruno Fosso[1], Vito Flavio Licciulli [3], Giorgio Grillo[3], Giacinto Donvito[4], Enrico Lavezzo[5], Giacomo Baruzzo [6], Graziano Pesole [1,2] and Monica Santamaria [1,*]

[1]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari 70126, Italy, [2]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari 'A. Moro', Bari 70126, Italy, [3]Institute of Biomedical Technologies, National Research Council, Bari Unit, Bari 70126, Italy, [4]National Institute for Nuclear Physics (INFN), Section of Bari, Bari 70126, Italy, [5]Department of Molecular Medicine, University of Padova, Padova 35131, Italy and [6]Department of Information Engineering, University of Padova, Padova 35131, Italy

*To whom correspondence should be addressed.
Associate Editor: Can Alkan

Received on February 25, 2021; revised on June 3, 2021; editorial decision on June 4, 2021; accepted on June 11, 2021

Abstract

Summary: ITSoneWB (ITSone WorkBench) is a Galaxy-based bioinformatic environment where comprehensive and high-quality reference data are connected with established pipelines and new tools in an automated and easy-to-use service targeted at global taxonomic analysis of eukaryotic communities based on Internal Transcribed Spacer 1 variants high-throughput sequencing.

Availability and implementation: ITSoneWB has been deployed on the INFN-Bari ReCaS cloud facility and is freely available on the web at http://itsonewb.cloud.ba.infn.it/galaxy.

Contact: m.santamaria@ibiom.cnr.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The amplicon targeted metagenomic analysis (here referred as DNA metabarcoding), in which taxon-related variants of selected genetic markers from environmental samples are explored through High-Throughput Sequencing (HTS) technologies, is largely applied to unravel the global composition of biotic communities in a rapid, massive and cheap way. After the widespread success of this approach in prokaryotic studies, a growing number of researchers interested in eukaryotic communities are now encouraged to use it. This results in an urgent need of increasingly comprehensive, well-controlled and FAIR-compliant (Wilkinson *et al.*, 2016) reference databases targeted to this domain, interfaced with established annotation tools and consistent taxonomies. Such a topic is becoming a primary target of some of the most important bioinformatics infrastructures as, for example, those established in Elixir and LifeWatch European ESFRI projects. In this framework, we developed the ITSone WorkBench (ITSoneWB), where comprehensive and high-quality Internal Transcribed Spacer 1 (ITS1) reference data, DNA metabarcoding well-established analysis pipelines and new tools are integrated in an easy-to-use service addressing the eukaryotic domain of life.

2 ITSoneWB rationale and novelty

A growing number of evidence have highlighted the great potential of ITS1 in discriminating Eukaryotes at deeper taxonomic levels, particularly in Fungi (Cheng *et al.*, 2016; Usyk *et al.*, 2017; Wang *et al.*, 2015). ITSoneWB provides the first integrated bioinformatic environment specifically targeted at this promising marker, where the user can easily submit his own sequences, access high-quality reference information and tools, and use them in customized and automated workflows without worrying about intermediate bioinformatics steps, often critical when data flow through different tools with different format requirements.

ITSoneDB (Santamaria *et al.*, 2018), to our knowledge the first and unique controlled and taxonomically referenced specialized collection of eukaryotic ITS1 sequences, is the core of the WorkBench. Last update, release 1.138 (March 2019), hosts 1 174 761 ITS1 sequences spanning 157 531 eukaryotic species. A section of the database, with 46 375 sequences belonging to 4115 species, is entirely dedicated to marine habitat. ITSoneDB has been developed in the framework of ELIXIR EXCELERATE project in order to enhance bioinformatic resources for metagenomic studies targeted to this particularly complex and still largely unexplored environment but

its usage actually embraces any eukaryotic sample. ITSoneDB reference dataset has been also recently made public in the ENA Browser under accession PRJEB33030.

Among the pipelines available to date in the WorkBench for sequence-based taxonomic assignment, users can choose BioMaS (Fosso *et al.*, 2015), QIIME (Caporaso *et al.*, 2010), QIIME2 (Bolyen *et al.*, 2019, 2) or Mothur (Schloss *et al.*, 2009). BioMaS, already freely available as a web-service at http://recasgateway.ba.infn.it/web/guest/biomas, offers an automated workflow for the taxonomic analysis of both prokaryotic and eukaryotic HTS DNA metabarcoding data. QIIME and QIIME2 are open-source pipelines for performing microbiome biodiversity analysis through quality graphics and statistics, and MOTHUR is a comprehensive suite of tools targeted at microbial community ecology. All these tools use ITSoneDB as reference database. Easy to use interfaces, available in the workbench, permit to execute the previously mentioned pipelines in an integrated environment (see supplementary material).

In addition, new services targeting some of the most common and challenging issues of metabarcoding experimental protocols, such as the design of effective universal primers and the evaluation of the barcoding gap in customized taxonomic ranges, are directly connected to ITSoneDB and accessible through WorkBench easy-to-use interfaces. The primers design tool aims at supporting researchers in designing successful 'universal' primer pairs able to amplify ITS1 in wide groups of organisms virtually avoiding, at the same time, any off-target amplification. This is still a tricky and crucial issue, since the use of ITS1 as taxonomic marker has gained popularity only recently (Badotti *et al.*, 2017; Usyk *et al.*, 2017; Wang *et al.*, 2015) and the already available primers pairs are often limited by taxonomic bias and able to generate only a low number of sequence reads, insufficient to encompass the global complexity of communities (Usyk *et al.*, 2017). We aimed to improve the primers inference by using the high-coverage ITSoneDB collection with Mopo16S (Sambo *et al.*, 2018), a recently developed primer inference tool. ITSoneWB allows to apply a modified version of Mopo16S to a set of ITS1 sequences extracted according to specific users' requests (e.g. a customized taxonomic target).

The barcoding gap estimation gives a prior idea of the ability of a specific genomic region to discriminate between taxa (Eckert *et al.*, 2014). The value of the barcoding gap, usually referred to the divergence between intra- and inter-specific sequence variability for congeneric DNA barcode sequences, strongly depends on taxonomic group and analytical practices (Čandek and Kuntner, 2015). Due to the important role of this parameter in predicting the experiment success, we developed and implemented in the WorkBench a new barcoding gap inference tool working on ITSoneDB collection. It can be applied among species belonging to the same genus or, at a higher taxonomic level, among genera belonging to the same family in customized taxonomic ranges.

Moreover, the user is provided with a facility, the ITSoneDB connector, to query, cross-referencing and downloading ITS1 data and metadata in case he wants to feed his own bioinformatic workflow.

Finally, in order to guarantee full interoperability with other Workflow Management Systems, we deployed a Dockerized version of ITSoneWB tools. Nonetheless, also a Dockerized version of the whole Galaxy environment is available. A complete documentation for both ITSoneWB and *ad hoc* developed tools installation and configuration is also available (itsonewb.readthedocs.io).

3 Conclusion

ITSoneWB is a new bioinformatic environment aimed at profiling community biodiversity based on ITS1, an increasingly popular DNA barcode in Eukaryotes. DNA metabarcoding established pipelines and new facilities are here oriented to ITSoneDB that hosts, in our knowledge, the first and unique specialized collection of well-controlled and taxonomically annotated ITS1 sequences embracing the entire Eukaryotic domain. The WorkBench is freely available and easy to use even by non-expert, and the executed analyses are easily reproducible in order to promote the data use and reuse according to the FAIR guidelines (Wilkinson *et al.*, 2016). Its virtual instance has been deployed on the ReCaS-Bari cloud facility thus supplying enough computational power and suitable scalability of the underlying resources in order to support large projects and/or to include new tools (Tangaro *et al.*, 2020).

Our next plan is to complete and increasingly enrich this virtual research environment by extending its application to additional eukaryotic taxonomic markers, allowing to use them individually or in combination, and enhancing the suite of accessory tools.

Acknowledgements

The authors thank Rob Finn (European Bioinformatics Institute—EMBL-EBI) and Nils Peder Willassen (UiT The Arctic University of Norway), for their support and cooperation in the recent development and future implementation of ITSoneDB in the framework of Elixir EXCELERATE project and Nicola Losito (Institute of Biomedical Technologies, National Research Council, Bari) for server management support.

Funding

This work was funded by ELIXIR, the research infrastructure for life-science data (ITSoneWB has been designed and developed in the mainframe of the ELIXIR 2017 Implementation Studies for integration of Italian node and the EXCELERATE marine Metagenomics use case), LifeWatch [Grant agreement ID: 211372], ELIXIR-IIB, ReCaS (Azione I—Interventi di rafforzamento strutturale, PONa3_00052, Avviso 254/Ric) and European Commission (ELIXIR-EXCELERATE HORIZON 2020). Funding for open access charge: ELIXIR-EXCELERATE HORIZON 2020 [grant agreement number 676559]. The Italian node is configured as a Joint Research Unit (JRU) named ELIXIR-IIB (formerly known as ELIXIR-ITA) within ELIXIR.

Conflict of Interest: none declared.

References

Badotti,F. *et al.* (2017) Effectiveness of ITS and sub-regions as DNA barcode markers for the identification of Basidiomycota (Fungi). *BMC Microbiol.*, **17**, 42.

Bolyen,E. *et al.* (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.*, **37**, 852–857.

Čandek,K. and Kuntner,M. (2015) DNA barcoding gap: reliable species identification over morphological and geographical scales. *Mol. Ecol. Resources*, **15**, 268–277.

Caporaso,J.G. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.

Cheng,T. *et al.* (2016) Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Mol. Ecol. Resources*, **16**, 138–149.

Eckert,E.M. *et al.* (2014) Does a barcoding gap exist in prokaryotes? Evidences from species delimitation in cyanobacteria. *Life (Basel)*, **5**, 50–64.

Fosso,B. *et al.* (2015) BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. *BMC Bioinformatics*, **16**, 203.

Sambo,F. *et al.* (2018) Optimizing PCR primers targeting the bacterial 16S ribosomal RNA gene. *BMC Bioinformatics*, **19**, 343.

Santamaria,M. *et al.* (2018) ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences. *Nucleic Acids Res.*, **46**, D127–D132.

Schloss,P.D. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Tangaro,M.A. *et al.* (2020) Laniakea: an open solution to provide Galaxy 'on-demand' instances over heterogeneous cloud infrastructures. *Gigascience*, **9**, giaa033.

Usyk,M. *et al.* (2017) Novel ITS1 fungal primers for characterization of the mycobiome. *mSphere*, **2**, e00488-17.

Wang,X.C. *et al.* (2015) ITS1: a DNA barcode better than ITS2 in eukaryotes? *Mol. Ecol. Resources*, **15**, 573–586.

Wilkinson,M.D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.