Published online 25 September 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D127–D132***
doi: 10.1093/nar/gkx855

ITSoneDB: a comprehensive collection of eukaryotic ribosomal RNA Internal Transcribed Spacer 1 (ITS1) sequences

Monica Santamaria[1,†], Bruno Fosso[1,†], Flavio Licciulli[2], Bachir Balech[1], Ilaria Larini[3], Giorgio Grillo[2], Giorgio De Caro[2], Sabino Liuni[2] and Graziano Pesole[1,3,*]

[1]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, Consiglio Nazionale delle Ricerche, Bari 70126, Italy, [2]Institute of Biomedical Technologies, Consiglio Nazionale delle Ricerche, Bari 70126, Italy and [3]Department of Biosciences, Biotechnology and Biopharmaceutics, University of Bari 'A. Moro', Bari 70126, Italy

Received August 08, 2017; Revised September 08, 2017; Editorial Decision September 12, 2017; Accepted September 18, 2017

ABSTRACT

A holistic understanding of environmental communities is the new challenge of metagenomics. Accordingly, the amplicon-based or metabarcoding approach, largely applied to investigate bacterial microbiomes, is moving to the eukaryotic world too. Indeed, the analysis of metabarcoding data may provide a comprehensive assessment of both bacterial and eukaryotic composition in a variety of environments, including human body. In this respect, whereas hypervariable regions of the 16S rRNA are the *de facto* standard barcode for bacteria, the Internal Transcribed Spacer 1 (ITS1) of ribosomal RNA gene cluster has shown a high potential in discriminating eukaryotes at deep taxonomic levels. As metabarcoding data analysis rely on the availability of a well-curated barcode reference resource, a comprehensive collection of ITS1 sequences supplied with robust taxonomies, is highly needed. To address this issue, we created ITSoneDB (available at http://itsonedb.cloud.ba.infn.it/) which in its current version hosts 985 240 ITS1 sequences spanning over 134 000 eukaryotic species. Each ITS1 is mapped on the NCBI reference taxonomy with its start and end positions precisely annotated. ITSoneDB has been developed in agreement to the FAIR guidelines by enabling the users to query and download its content through a simple web-interface and access relevant metadata by cross-linking to European Nucleotide Archive.

INTRODUCTION

The increasing amount of DNA sequence information generated by worldwide metagenomics initiatives is enabling the identification of a growing number of taxa and genes in any natural and anthropic environment. Most of microbiome studies so far have been focused on the assessment of the taxonomic composition of prokaryotic communities based on one or more hypervariable regions of the 16S rRNA. Nowadays, a more holistic investigation, including also viral and eukaryotic components, is becoming the new challenge (1). In this framework, the amplicon-based metagenomic approach or metabarcoding, in which selected short, variable and standardized DNA regions, named DNA barcodes (2,3), are simultaneously amplified and sequenced from an ensemble of organisms sharing the same habitat, is rapidly gaining popularity also to unravel eukaryotic diversity. In addition to Fungi, which are particularly intriguing due to their ubiquity, high diversity and often cryptic manifestations, scientists are looking with growing interest at other Eukaryotes (4–7). Indeed, despite their widespread, diversity and their important role in biogeochemical cycles (8–11), eukaryotic microbial species appear to be consistently underestimated. This has been recently confirmed in the marine environment by Tara Oceans Initiative (12). Finally, metabarcoding has been used also to monitor the macro-fauna biodiversity in aquatic environments (13).

The taxonomic classification of high-throughput sequence reads generated by metabarcoding experiments is usually based on first mapping each read to the relevant barcode reference collection and then inferring its more likely taxonomic attribute at species or higher taxonomic rank level (14). In this respect, reference databases aimed at supporting eukaryotic communities characterization are still far to be reliable and exhaustive, possibly causing biased taxonomic inferences. In order to address this drawback

*To whom correspondence should be addressed. Tel: +39 080 544 3588; Fax: +39 080 544 3317; Email: graziano.pesole@uniba.it
†These authors contributed equally to the paper as first authors.

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

a worldwide remarkable effort is on-going to collect and harmonize the vast amount of sequence data already available for several barcodes, including the Internal Transcribed Spacer (ITS) of the rRNA gene cluster, one of the most promising for the assessment of eukaryotic biodiversity. ITS has been already proposed as the standard DNA barcode for fungi and plants (15–18) and widely used for discriminating taxa in other biological groups, including algae, protists and animals (4,19,20). In particular, many recent lines of evidence have highlighted the great potential of the ITS1 sub-region in discriminating eukaryotes at deeper taxonomic levels, particularly in Fungi. A comparative analysis between ITS1 and ITS2 in 10 major groups of eukaryotes, in terms of PCR primer universality, length of amplification product and GC content affecting DNA sequencing outcome and species discrimination power, robustly supports the hypothesis that ITS1 is a better DNA barcode than ITS2 for eukaryotic species (4). However, the availability of comprehensive and quality-controlled resources of ITS1 sequences, supplied with unbiased and unambiguous taxonomies and interfaced with the state of the art pipelines for metagenomics analysis, is still lacking. Such resources should be taxonomically comprehensive and unbiased, well controlled and their content easily accessible and retrievable. This is a critical obstacle to take full advantage from the most reliable bioinformatic tools, such as QIIME (21) and its recent upgrade QIIME 2 (https://qiime2.org/), Mothur (22), MICCA (23) or BioMaS (Bioinformatic analysis of metagenomics ampliconS) (24), that rely on comparing the meta-barcode sequences against the relevant reference databases to infer their taxonomic origin.

In order to address this gap we have developed ITSoneDB (http://itsonedb.cloud.ba.infn.it/), focusing on the whole eukaryotic domain and establishing the first and unique curated reference database aimed at ITS1-based metagenomic surveys. Indeed, similar well-annotated and updated databases, such as UNITE (User-friendly Nordic ITS Ectomycorrhiza Database, http://unite.ut.ee) and ITS2 Database (http://its2.bioapps.biozentrum.uni-wuerzburg.de/) concern either the entire ITS sequence or its ITS2 sub-region, respectively. Currently, ITSoneDB collects about one million ITS1 sequences spanning over 134 000 species (according to NCBI taxonomy). The annotation of ITS1 region boundaries has been refined by coupling the information available in the original European Nucleotide Archive (ENA) entries with those inferred by mapping Hidden Markov Models (HMM) corresponding to the conserved ITS1 flanking genes (see 'Materials and Methods' section).

We have undertaken a number of actions to ensure that our data follow the FAIR principles (25). In particular, Findability and Accessibility are already granted by user-friendly query and cross-link systems to retrieve and download the sequences stored in the database and get their associated metadata respectively. We are working to extend the Interoperability and the Re-usability features by integrating ITSoneDB in a cloud-based Galaxy workbench in which users may run established metagenomics analysis pipelines, thus providing complete and reusable workflows for taxonomic annotation of eukaryotic microbiomes. ITSoneDB can be easily interfaced with state of the art metabarcoding analysis tools such as QIIME (21) or BioMaS (24) as well as other popular metagenomic analyses pipelines. Moreover, we plan to integrate our database in the EBI metagenomics portal in order to increase its accessibility and use.

DATABASE CONTENT

Currently, ITSoneDB collects 985 240 ITS1 sequences corresponding to 134 598 species (according to NCBI taxonomy), and 276 362 (28%) ITS1 region positions in the original sequences are inferred only by HMM profiles mapping while 543 266 (55.2%) are obtained from the ENA entries features table. The location of ITS1 region in 165 612 (16.8%) sequences are inferred by both the approaches. Table 1 reports the number of sequences and species in the eukaryotic kingdom and in its major taxa represented in ITSoneDB, where Fungi are the most represented taxonomic groups, covering almost 70% of database content. Supplementary Figure S1 shows a more detailed taxonomic spread of ITSoneDB sequences across Eukaryotes. The length of ITS1 sequences collected in ITSoneDB mainly ranges between 50 and 1000 bp, with 91.7% of the sequence between 100 and 300 bp long.

Each ITSoneDB entry is composed of three main sections: the first consists of an overall entry description in which general information about the entire sequence, such as coverage, function and length, and taxonomic classification are reported. The second one, named 'ITS1 sequence', reports information about the ITS1 region annotation inferred from ENA feature tables and/or HMM profiles mapping. The last section, indicated as '18S rRNA HMM profile — target sequence alignment' and '5.8S rRNA HMM profile — target sequence alignment' (see Figure 1), only available if ITS1 boundaries have been refined by HMM mapping, shows the alignments of HMM profiles and the corresponding regions in the sequence.

A regular update of ITSoneDB is planned on a six-month based interval according to the most recent ENA release.

DATABASE FEATURES

ITSoneDB is publicly and freely accessible through web browser on a permanent URL. Single or multiple entries can be selected for web visualization and/or retrieval through different query options (located at the top left of the home page). The 'simple search' box allows querying the database by species name/s, taxon name/s or ENA accession number/s. An auto-completion feature permits to choose easily the desired query terms. The 'tree search' option allows a simple navigation across the taxonomic tree (NCBI taxonomy) and the selection of the taxa of interest by checking the adjacent box; the total number of ITS1 sequences with ENA and HMM localization are displayed next to taxon name. Alternatively, the 'advanced search' option allows constructing a refined query using boolean operators on the queries performed previously (shown in the 'executed query' box). For instance, prior queries for the species *Aspergillus aculeatus* (query#1) and *Zygowilliopsis californica* (query#2) performed separately, can be combined to obtain all entries of the two species through a composing panel as follows: query#1 OR query#2. The advanced search may be also refined by defining parameters

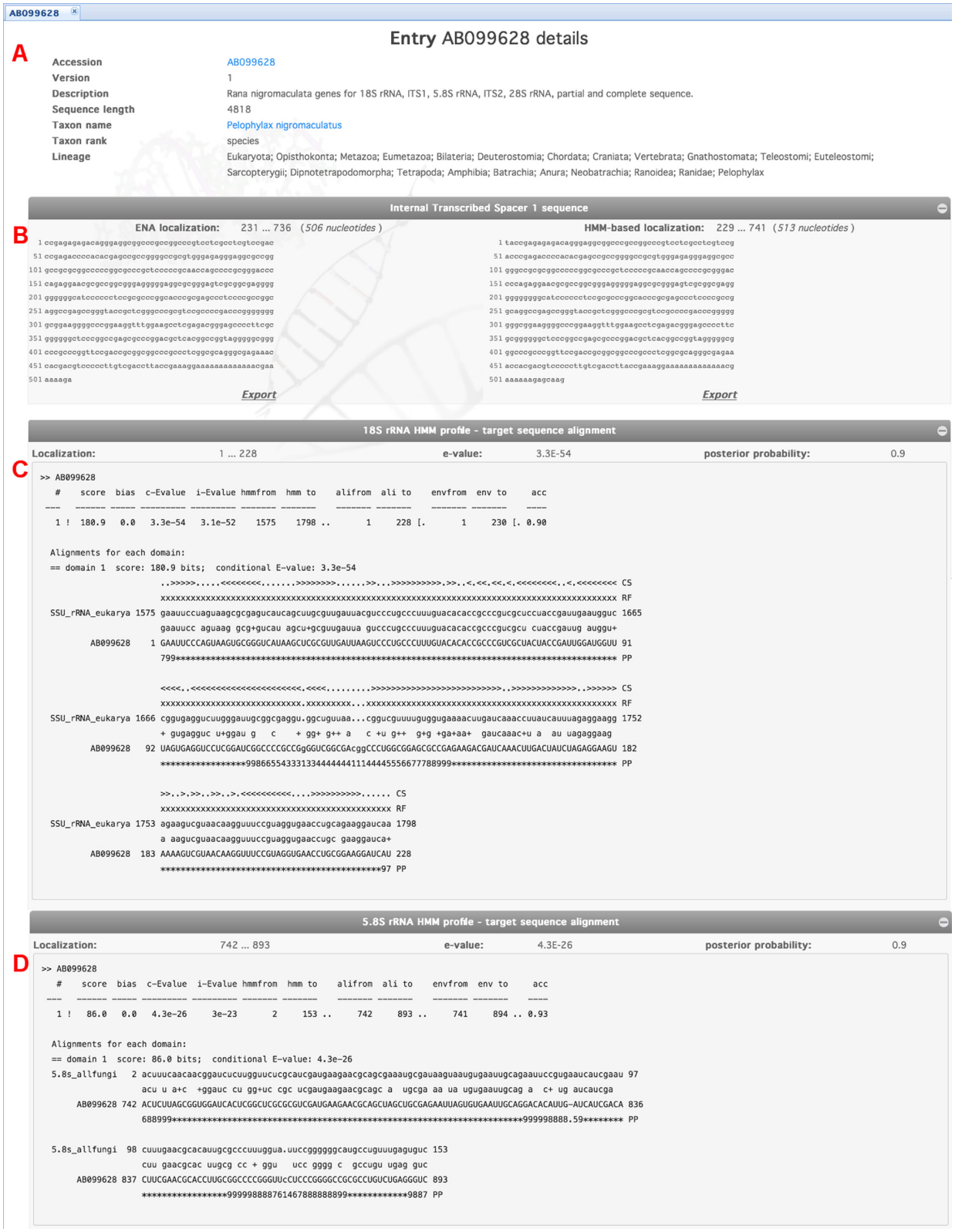

Figure 1. Snapshot of a ITSoneDB entry. In (**A**) the entry information directly extracted from ENA are reported: (i) accession number (with a hyperlink to the corresponding ENA item), (ii) version, (iii) description, (iv) sequence length (the whole sequence length, not only the ITS1), (v) taxon name (scientific name of the organism the sequence belongs, with a hyperlink to the corresponding NCBI taxonomy entry), (vi) taxon rank (taxonomic class) and lineage (full taxonomic path associated to the Taxon name). In (**B**) the ITS1 position annotation from ENA and HMM are reported. (**C**) and (**D**) show the alignments of the sequence with 18S and 5.8S rRNA HMM profiles respectively.

Table 1. ITSoneDB content statistics

Taxon	Taxid	Total sequences	ENA annotation only	HMM annotation only	ENA and HMM	Species
Eukaryota	2759	985 240	543 266	276 362	165 612	134 598
Fungi	4751	684 540	378 049	221 723	84 768	53 552
Metazoa	33 208	54 782	32 186	9084	13 512	9438
Viridiplantae	33 090	203 437	113 572	32 503	57 362	66 595

regarding sequence length, ITS1 annotation method (ENA, HMM or both) and the *E*-values and/or posterior probability values supporting the HMM matches.

Each entry in the query output can be visualized as a web page showing the accession number (cross-linked to ENA), a brief description, the full lineage description from the NCBI taxonomy (hyper-linked to the NCBI taxonomy database), the sequence length and taxon rank. The entries of interest can be exported into FASTA formatted DNA sequence file by choosing one or both annotation options (ENA annotations and/or HMM). Moreover, ITSoneDB offers an additional export feature limited to a representative sequence per species that returns the centroid of a population of sequences belonging to the same species (see 'Materials and Methods' section for additional details). Another important feature of ITSoneDB, especially for local analysis, is the possibility to export the entire database or the species representative sequences (options available at top right of the home page).

MATERIALS AND METHODS

In order to generate, maintain and update ITSoneDB, we designed a multi-step Python and BASH workflow (Supplementary Figure S2). In the first step, the nucleotide entries from the ENA (European Nucleotide Archive, http://www.ebi.ac.uk/ena) database (26) are locally downloaded. The current version of ITSoneDB 1.131, which extends and upgrade a previous version limited to Fungi (27), has been populated by considering the ENA release 131 (02/27/2017), counting for 803 147 518 entries.

In order to reduce both the computational requirements and processing time, the Plant (PLN), Environmental (ENV), Human (HUM), Fungal (FUN), Other Mammal (MAM), Invertebrates (INV), Other Vertebrates (VRT), *Mus musculus* (MUS), Other Rodents (ROD), Unclassified (UNC) divisions belonging to the Genome Sequencing Scan (GSS), High-Throughput cDNA Sequencing (HTC), High-Throughput Genome Sequencing (HTG), Standard (STD), Patent (PAT), EST (expressed sequence tag) and Transcriptome Shotgun Assembly (TSA) data classes, have been locally downloaded. Afterward, only the Eukaryotic entries have been selected, reducing the data count to 83 008 723 items.

In the second step, the accession number, the description and the available annotation under specific feature keys (i.e. rRNA, misc_rRNA, misc_feature and source) have been extracted from each entry and stored in a TSV (tab-separated values) file, while the associated nucleotide sequence has been saved in FASTA file. This procedure allowed also to associate taxonomic information (i.e. NCBI taxonomy identifier and taxonomic path) to each ENA accession number. Subsequently, a comprehensive, *ad hoc* developed and manually curated dictionary of 110 common ITS1 synonyms (see Supplementary Table S1) has been used to filter the data stored in the TSV files and select the entries where the ITS1 start and end positions were specifically annotated.

At the same time, the ITS1 boundaries have been validated or *de novo* defined by using a similarity-based approach. ITS1 is flanked by two highly conserved genes encoding for the ribosomal RNA 18S and 5.85, respectively, whose conservation profile can be suitably modeled by HMM. The HMMs for 18S (RF01960) and 5.8S (RF00002) rRNAs have been generated by using the reference multiple alignments (Stockholm format) available in the RFAM database (28,29). The 18S and 5.8S rRNA HMMs have been thus mapped against previously extracted ENA sequences by using *hmmsearch* (HMMER 3.1 package) (30). Statistically significant HMM matches have been considered for ITS1 boundaries definition using as threshold the e-value ≤ 0.001. In order to retain matches where the terminal portion of the 18S HMM profile and/or the initial portion of the 5.8S HMM profile aligned to initial or terminal part of the query sequence, respectively, we also considered matches with *E*-value ≥ 0.001 but with an average posterior-probability (a measure of alignment accuracy) >0.85 (30). The information regarding ITS1 boundaries extracted from entry annotation and/or defined by inferring the 18S and 5.8S locations were then merged to generate the tables used to populate the database.

Finally, for each species represented in the database a representative entry was selected: all the sequences belonging to the same species were extracted and clustered, setting up a 97% identity threshold, by applying VSEARCH (31). The reference sequence corresponded the centroid of the largest cluster.

DATABASE ARCHITECTURE AND WEB INTERFACE

ITSoneDB implementation is based on a three-tier architecture: client, server and database (Supplementary Figure S3). In the database layer, data and metadata are stored in a MySQL (version 5.5) relational DBMS (Database Management System) using INNODB as stored engine in order to implement persistent queries. The Graphical User Interface (GUI) is developed as JAVA Web Application in Java Platform Enterprise Edition — Java EE. It uses jQuery/jQuery-UI framework JavaScript on the client layer, Java servlets and JavaServer Pages (jsp) on the server layer. The web application is deployed in a Tomcat web server (https://tomcat.apache.org).

To implement the communication between the data layer and the Web Application, the Hibernate ORM (Object Relational Mapping, http://hibernate.org/orm/) has been adopted. It provides a framework for mapping an object-

oriented domain model to a relational database enabling us to handle the data layer as objects in the web pages.

FUTURE DEVELOPMENTS

The first release of ITSoneDB, together with all the tools developed to build and populate it, has been designed to be further improved and expanded. Many of the future improvements will be carried out in the framework of the activities of the ELIXIR research infrastructure for biological data (https://www.elixir-europe.org) in collaboration with the EMBL-EBI and Norway nodes. First, we plan to constantly update the data by retrieving and curating the new ITS1 sequences available in primary databases and other resources, including those hosting shot-gun and amplicon-based metagenomics datasets. During these updates, we will take all necessary actions to maintain the long-time usability and value of our data by keeping in mind the FAIR data principles guidelines. We plan also to connect ITSoneDB to the UniEuk Initiative (32) (http://unieuk.org) in order to map its content on a more curated and harmonized taxonomy. In order to further support users involved in metagenomics experiments we plan to implement services for (i) calculating the 'barcoding gap' in a custom-defined taxonomic range; (ii) designing 'universal' PCR primers effective in that range; and (iii) performing multi-query sequence similarity searches suitable for large metagenomic datasets. A new section of the database, will be also created, focused on organisms living in the marine environment in order to address its largely unknown complexity. This section will be linked to other Marine reference databases such as MarRef, MarDB and MarCat (available at https://mmp.sfb.uit.no) developed within the ELIXIR project framework.

We will allow to efficiently parallelize and manage the pairwise and multiple alignments required by the new planned functions (barcoding gap computation and primers design respectively) by integrating ITSoneDB in a Cloud-based analysis workspace. Furthermore, we will release and update *ad hoc* pre-formatted version of ITSoneDB sequences and taxonomy in order to allow the use of popular metabarcoding analysis tools such as BioMaS (24), QIIME (21), LCA-classifier (used by META-pipe, https://arxiv.org/abs/1604.04103), MAPseq (used by EBI metagenomics—EMG (33)) (http://www.biorxiv.org/content/early/2017/04/12/126953), MOTHUR (22) or UCHIME (34). Sequences, taxonomy and other metadata will be also made available in human and machine readable tabular formats. We also plan to create a Galaxy workbench in order to allow users to select a specific analysis pipeline for metabarcoding data analysis using ITSoneDB as a reference collection, fostering data sharing, transparency and reproducibility. Finally, we will work in collaboration with EBI in the Elixir framework to provide ITSoneDB as a reference database in a specific workflow for ITS1, within the EBI metagenomics portal. This will further increase its use, exposure and interoperability.

DATA AVAILABILITY

ITSoneDB is freely accessible as a web application at http://itsonedb.cloud.ba.infn.it/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We would like to thank Rob Finn (European Bioinformatics Institute - EMBL-EBI) and Nils Peder Willassen (UiT The Arctic University of Norway), Giacinto Donvito (National Institute of Nuclear Physics - INFN) and Giorgio Pietro Maggi (INFN and Politecnico di Bari) for their support and cooperation in the recent development and future implementation of ITSoneDB in the framework of Elixir EXCELERATE project, Nicola Losito (Institute of Biomedical Technologies, National Research Council, Bari) for server management support.

FUNDING

Italian Ministry for Education, University and Research; LifeWatch, ELIXIR-IIB, 'PON Ricerca e Competitività 2007–2013' Program; ReCaS (Azione I—Interventi di rafforzamento strutturale, PONa3_00052, Avviso 254/Ric); PRISMA (Asse II—Sostegno all'innovazione, PON04a2_A); European Commission (ELIXIR-EXCELERATE HORIZON 2020, EMBRIC HORIZON 2020). Funding for open access charge: ELIXIR-EXCELERATE HORIZON 2020 [GA 67559].
Conflict of interest statement. None declared.

REFERENCES

1. Uyaguari-Diaz,M.I., Chan,M., Chaban,B.L., Croxen,M.A., Finke,J.F., Hill,J.E., Peabody,M.A., Van Rossum,T., Suttle,C.A., Brinkman,F.S. *et al.* (2016) A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome*, **4**, 20.
2. Hebert,P.D. and Gregory,T.R. (2005) The promise of DNA barcoding for taxonomy. *Syst. Biol.*, **54**, 852–859.
3. Hebert,P.D.N., Cywinska,A., Ball,S.L. and DeWaard,J.R. (2003) Biological identifications through DNA barcodes. *Proc. R. Soc. B Biol. Sci.*, **270**, 313–321.
4. Wang,X.C., Liu,C., Huang,L., Bengtsson-Palme,J., Chen,H.M., Zhang,J.H., Cai,D.Y. and Li,J.Q. (2015) ITS1: a DNA barcode better than ITS2 in eukaryotes? *Mol. Ecol. Resour.*, **15**, 573–586.
5. Pernice,M.C., Giner,C.R., Logares,R., Perera-Bel,J., Acinas,S.G., Duarte,C.M., Gasol,J.M. and Massana,R. (2016) Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME J.*, **10**, 945–958.
6. Hugon,P., Lagier,J.C., Colson,P., Bittar,F. and Raoult,D. (2017) Repertoire of human gut microbes. *Microb. Pathog.*, **106**, 103–112.
7. Rusin,L.Y. (2016) [Metagenomics and biodiversity of sphagnum bogs]. *Mol. Biol. (Mosk)*, **50**, 730–734.
8. Debroas,D., Domaizon,I., Humbert,J.F., Jardillier,L., Lepere,C., Oudart,A. and Taib,N. (2017) Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data. *FEMS Microbiol. Ecol.*, **93**, doi:10.1093/femsec/fix023.
9. Caron,D.A., Worden,A.Z., Countway,P.D., Demir,E. and Heidelberg,K.B. (2009) Protists are microbes too: a perspective. *ISME J.*, **3**, 4–12.
10. Grattepanche,J.D., Santoferrara,L.F., McManus,G.B. and Katz,L.A. (2014) Diversity of diversity: conceptual and methodological differences in biodiversity estimates of eukaryotic microbes as compared to bacteria. *Trends Microbiol.*, **22**, 432–437.
11. Sherr,E. and Sherr,B. (1988) Role of microbes in pelagic food webs—a revised concept. *Limnol. Oceanogr.*, **33**, 1225–1227.

12. de Vargas,C., Audic,S., Henry,N., Decelle,J., Mahe,F., Logares,R., Lara,E., Berney,C., Le Bescot,N., Probert,I. *et al.* (2015) Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.
13. Yamamoto,S., Masuda,R., Sato,Y., Sado,T., Araki,H., Kondoh,M., Minamoto,T. and Miya,M. (2017) Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Sci. Rep.*, **7**, 40368.
14. Droge,J. and McHardy,A.C. (2012) Taxonomic binning of metagenome samples generated by next-generation sequencing technologies. *Brief. Bioinform.*, **13**, 646–655.
15. Xu,J. (2016) Fungal DNA barcoding. *Genome*, **59**, 913–932.
16. Cheng,T., Xu,C., Lei,L., Li,C., Zhang,Y. and Zhou,S. (2016) Barcoding the kingdom Plantae: new PCR primers for ITS regions of plants with improved universality and specificity. *Mol. Ecol. Resour.*, **16**, 138–149.
17. Sass,C., Little,D.P., Stevenson,D.W. and Specht,C.D. (2007) DNA barcoding in the cycadales: testing the potential of proposed barcoding markers for species identification of cycads. *PLoS One*, **2**, e1154.
18. Giudicelli,G.C., Mader,G. and de Freitas,L.B. (2015) Efficiency of ITS sequences for DNA barcoding in Passiflora (Passifloraceae). *Int. J. Mol. Sci.*, **16**, 7289–7303.
19. Hadi,S.I., Santana,H., Brunale,P.P., Gomes,T.G., Oliveira,M.D., Matthiensen,A., Oliveira,M.E., Silva,F.C. and Brasil,B.S. (2016) DNA barcoding green microalgae isolated from neotropical inland waters. *PLoS One*, **11**, e0149284.
20. Perez-del-Olmo,A., Georgieva,S., Pula,H.J. and Kostadinova,A. (2014) Molecular and morphological evidence for three species of Diplostomum (Digenea: Diplostomidae), parasites of fishes and fish-eating birds in Spain. *Parasit. Vectors*, **7**, 502.
21. Caporaso,J.G., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,F.D., Costello,E.K., Fierer,N., Pena,A.G., Goodrich,J.K., Gordon,J.I. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.
22. Schloss,P.D., Westcott,S.L., Ryabin,T., Hall,J.R., Hartmann,M., Hollister,E.B., Lesniewski,R.A., Oakley,B.B., Parks,D.H., Robinson,C.J. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.
23. Albanese,D., Fontana,P., De Filippo,C., Cavalieri,D. and Donati,C. (2015) MICCA: a complete and accurate software for taxonomic profiling of metagenomic data. *Sci. Rep.*, **5**, 9743.
24. Fosso,B., Santamaria,M., Marzano,M., Alonso-Alemany,D., Valiente,G., Donvito,G., Monaco,A., Notarangelo,P. and Pesole,G. (2015) BioMaS: a modular pipeline for Bioinformatic analysis of Metagenomic AmpliconS. *BMC Bioinformatics*, **16**, 203.
25. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.W., da Silva Santos,L.B., Bourne,P.E. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
26. Toribio,A.L., Alako,B., Amid,C., Cerdeno-Tarraga,A., Clarke,L., Cleland,I., Fairley,S., Gibson,R., Goodgame,N., Ten Hoopen,P. *et al.* (2017) European nucleotide archive in 2016. *Nucleic Acids Res.*, **45**, D32–D36.
27. Santamaria,M., Fosso,B., Consiglio,A., De Caro,G., Grillo,G., Licciulli,F., Liuni,S., Marzano,M., Alonso-Alemany,D., Valiente,G. *et al.* (2012) Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform.*, **13**, 682–695.
28. Daub,J., Eberhardt,R.Y., Tate,J.G. and Burge,S.W. (2015) Rfam: annotating families of non-coding RNA sequences. *Methods Mol. Biol.*, **1269**, 349–363.
29. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
30. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
31. Rognes,T., Flouri,T., Nichols,B., Quince,C. and Mahe,F. (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ*, **4**, e2584.
32. Berney,C., Ciuprina,A., Bender,S., Brodie,J., Edgcomb,V., Kim,E., Rajan,J., Parfrey,L.W., Adl,S., Audic,S. *et al.* (2017) UniEuk: time to speak a common language in protistology! *J. Eukaryot. Microbiol.*, **64**, 407–411.
33. Mitchell,A., Bucchini,F., Cochrane,G., Denise,H., ten Hoopen,P., Fraser,M., Pesseat,S., Potter,S., Scheremetjew,M., Sterk,P. *et al.* (2016) EBI metagenomics in 2016–an expanding and evolving resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res.*, **44**, D595–D603.
34. Edgar,R.C., Haas,B.J., Clemente,J.C., Quince,C. and Knight,R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194–2200.