UniEuk

Universal taxonomic framework and integrated reference gene databases for Eukaryotic biology, ecology, and evolution

November 8, 2017

Dear Colleagues,

*Uni**Euk*** is an open, inclusive, community-based and expert-driven international initiative to build a flexible, adaptive universal taxonomic framework for eukaryotes. It provides an online environment and simple tools to unite community knowledge with morphological and genetic data on protist diversity. The UniEuk taxonomy will be implemented at EMBL-EBI, ENA and ultimately be included in the NCBI taxonomy database, ensuring its long-term preservation and universal access for science and education. *For details, please see attached paper and Figure 1 below*.

One of the three UniEuk modules --*EukBank* (Fig. 1)-- will facilitate and standardize the analysis of high-throughput DNA/RNA metabarcoding (HTM) surveys that are now being carried out in many studies, and enable the incorporation of this HTM data into the *UniEuk* system. Combining an ultra-fast algorithm generating **stable** clusters of amplicons (*Swarm*, Mahé et al. 2015) and state-of-the-art methods of phylogenetic placement (*EPA*-based; Berger et al. 2011), the *EukBank* will combine all eukaryotic HTM datasets into a single homogenous database, and allow sorting and phylogenetic placement of the novel diversity into community agreed reference trees (from *EukRef* in Fig. 1). The aim is to **standardize** observations **of global eukaryotic diversity across biomes** (e.g., saturation, relative frequencies, phylogeny), and allow **identification and preliminary naming of novel eukaryotic lineages of ecological and/or phylogenetic relevance**. These will inform the *UniEuk* taxonomic framework (*EukMap* in Fig. 1), thus highlighting eukaryotic groups that warrant further investigation.

EukBank v1.0 will begin by curating **18S V4 rRNA metabarcoding datasets** to test its robustness and scalability. Subsequent updates will include datasets derived from other primer sets. **We need your help to successfully launch *EukBank* v1.0 in the next 9 months**, and we are **seeking your contribution of any published or unpublished V4 datasets (in FASTQ format), along with a minimum set of metadata** (MiMARKs standards)**, by the end of the year (December 31st, 2017)**. These data will help to screen and eliminate problems moving forward. We have started with >650 V4-sequenced samples from diverse biomes (marine sediments and plankton, abyssal waters, tropical and mountain forests soils, fresh water), and preliminary analyses are exciting (Fig. 2).

Data providers who would like to participate in the first EukBank v1.0 community paper will contribute their published or unpublished data through the **European Nucleotide Archive (EMBL-EBI, ENA) platform**. With the provider's consent, the ENA team will share the data with UniEuk/EukBank team through a **private password protected** server, so that downstream analysis can be performed. Providers of unpublished data can choose to release their data publically at any time, but no latter than the publication date of the community paper (in 2018). **Owners of datasets already deposited at NCBI and still under embargo are encouraged to release their data in order for data inclusion in the EMBL-EBI platform. In this case, please email datasubs@ebi.ac.uk to inform us of the release. If embargo is an issue, please contact us directly to discuss alternative routes into EukBank. In any cases, please always use 'UniEuk_EukBank' in the subject line of your email.**

For more details on how to share your data with the EukBank team and how to specify a release date for unpublished data, please see **'UniEuk_EukBank_submission_guide'** attached; and email **datasubs@ebi.ac.uk**, should you need further clarifications.

Partners

On the short term, you will receive:

- **co-authorship on a high-impact, community paper launching the *EukBank* and showing its power** to explore global protist biodiversity patterns (saturation across phylogenetic and spatial scales, *alpha* and *beta*-diversity across biomes, new lineages, etc). Co-authorship is offered to the Principle Investigator and one designate for each contributed published or unpublished V4 dataset.
- a streamlined protocol and help from the ENA team to submit past and current HTM datasets to EBI-ENA, and get their Accession # (see attached **submission guide**).
- a table of taxonomically annotated OTUs generated from all the datasets used, which would enable you to explore your dataset within a global, standardized context of protist diversity.
- a simple protocol and state-of-the-art tools (EPA-ng /next generation) to place your OTUs into the **UniEuk synthetic Tree of Eukaryotic Life** and any sub-group trees your may have, and thus explore the phylogenetic richness, abundance, and diversity of your data in a standardized framework.

In the longer term, the permanent implementation of the EukBank at EMBL-EBI, ENA will ensure sustainability and offer the following:

- a simple way to submit and get Accession # for eukaryotic metabarcoding datasets (multiple markers).
- a standardized public database of raw and clustered environmental reads for assessing protist diversity, with user-friendly tools to explore their novelty and ecology. Note that the fast-growing INSDC sequence-read (short-read) archive is already 5,000 times larger than the INSDC sequence domain (assembled and annotated sequences), which makes it impossible to use classical tools such as BLAST to explore the data. **By pooling and reducing the complexity of eukaryotic HTM datasets into a unique and community-accessible reference repository, the EukBank will empower scientists to easily submit, explore (e.g. via *Blast*), visualize and retrieve protist diversity data.**
- a solution to quickly detect novel eukaryotic lineages of ecological and/or evolutionary relevance, and inject them into the UniEuk taxonomic framework.

We thank you in advance for your participation to this collective effort toward a growing global perspective on eukaryotic diversity, and for spreading the news to any colleagues who might be interested. We look forward to working with you to release EukBank v1.0.

With Our Best Regards,
Members of the *UniEuk* Committees, (http://unieuk.org/people/)

UniEuk

Universal taxonomic framework and integrated reference gene databases for Eukaryotic biology, ecology, and evolution

DATA SUBMISSION: see the **'UniEuk_EukBank_submission_guide'** attached.
In summary, all you need to do is to fill-up **a table**, with minimum metadata information.
Up to several hundred samples can be submitted in bulk.
Information attached to each sample is:

- *taxonomic name*:	**"uncultured eukaryote"**
- *collection date*:	you know when it happened
- *geographic location*:	country and/or sea, you will be guided through an official list
- *latitude longitude*	(if applicable)
- *environment (biome)*:	www.environmentontology.org/Browse-EnvO
- *environment (feature)*:	www.environmentontology.org/Browse-EnvO
- *environment (material)*:	www.environmentontology.org/Browse-EnvO
- *target_gene*:	**18S SSU rRNA**
- *target_subfragment*:	**V4**
- *pcr_primers*:	names and sequences
- *instrument platform/model*:	**Illumina MiSeq**

There will also be optional fields available to enrich the sample annotation, including *depth/altitude*, *organismal size fraction*, *user-defined fields*; the more annotation you can provide for your sample the better as it adds context.

All experimental information is recorded during the '*experiment*' step.

Submission can be done manually **or in bulk**, **and we can assist you**. Please write to: **datasubs@ebi.ac.uk** and add « **UniEuk_EukBank** » in the subject line of your email.

DATA RELEASE POLICY: When submitting a raw dataset to EMBL-EBI, ENA, a user participating in the *EukBank* project can choose between "*immediate*" or "*delayed*" publication. Publicly available datasets are immediately processed and included in the next EukBank release, which will occur every 6 weeks. Datasets under '*delayed publication*' are not public but incorporated into the *EukBank* for immediate release upon author's decision or publication of the data.

UniEuk

***Uni**versal taxonomic framework and integrated reference gene databases for **Euk**aryotic biology, ecology, and evolution*

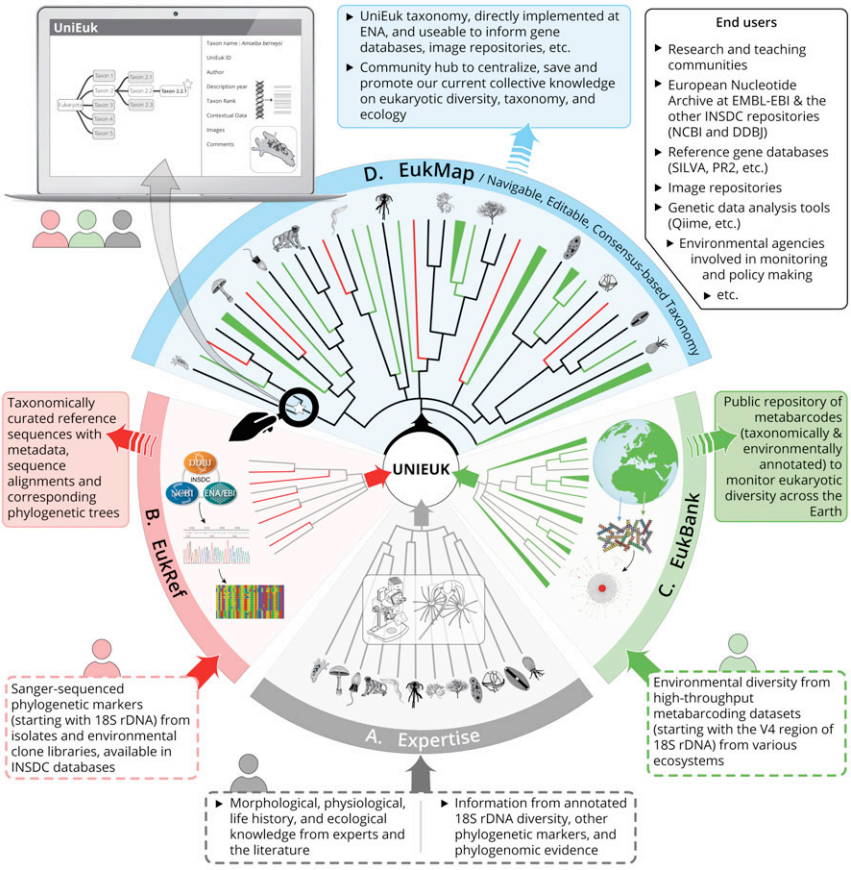

Figure 1: The *UniEuk* information and work-flow. Bottom-up, community-based information on eukaryotic biodiversity from (A) classical knowledge, (B) phylogenetic diversity, and (C) environmental '-omics' surveys (the ***EukBank***!), converge and synergize through the *UniEuk* modules to inform the navigable and editable, consensus-based taxonomic framework (D). Dotted and colored frames indicate input and output information, respectively.

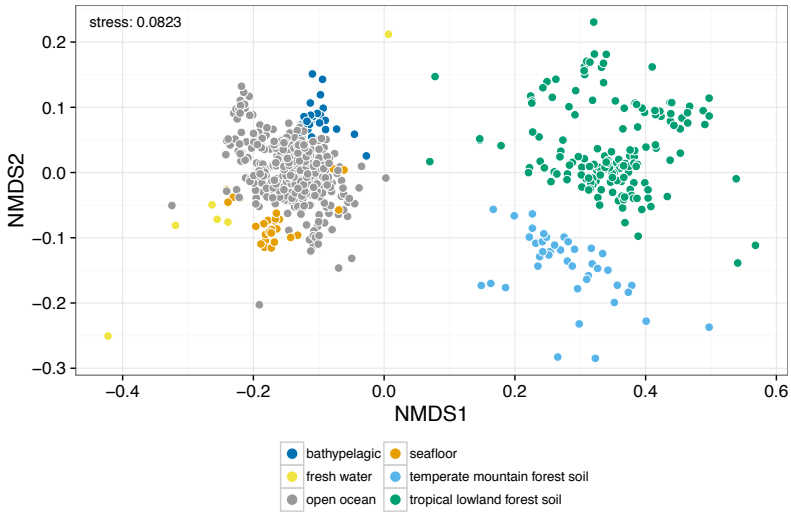

Figure 2: Organizing and exploring total eukaryotic diversity. Preliminary meta-analysis of eukaryotic community similarity (V4 rDNA/RNA data) from 672 samples from marine sediments and plankton, abyssal waters, soils from tropical and mountain forests, fresh water, etc. Note the overall structuring of communities by biomes, and the much greater community differentiation in terrestrial ecosystems, as compared to the world oceans.

Partners