

UniEuk

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

=====

Steering Committee (SC)

Colomban de Vargas

France, CNRS - Station Biologique de Roscoff, Roscoff – BioMarKs, Tara-Oceans, & CBOL ProWG, coordinator / UniEuk director

Pelin Yilmaz

Germany, Max Planck Institute & Jacobs University, Bremen – chair of SILVA's eukaryote taxonomy working group (ETWG) / UniEuk co-director

Sina Adl

Canada, University of Saskatchewan, Saskatoon – leader of ISOP's universal eukaryotic taxonomic framework

Juliet Brodie

UK, The Natural History Museum, London – board member of the International Phycological Society

Guy Cochrane

UK, EMBL-EBI / ENA, Cambridge – head of the European Nucleotide Archive at EMBL-EBI

Javier del Campo

Canada, University of British Columbia, Vancouver – EukRef coordinator

Virginia Edgcomb

USA, Woods Hole Oceanographic Institution, Woods Hole MA – board member of the International Society of Protistologists

Stefan Geisen

The Netherlands, Netherlands Institute of Ecology, Wageningen – leader of the Soil Protists Initiative

Frank Oliver Glöckner

Germany, Max Planck Institute & Jacobs University, Bremen – head of SILVA and MicroB3/OSD

Eunsoo Kim

USA, American Museum of Natural History, New York City NY – curator of Protistology at the AMNH

Alastair Simpson

Canada, Dalhousie University, Halifax – chair of ISOP's protist systematics committee

Advisory Council (AC) - as of May 2016

David Caron (AC chair)

Sandra Baldauf

Sonya Dyhrman

Laura Katz

Connie Lovejoy

Alexandra Worden

John Archibald

David Bass

Patrick Keeling

Jan Pawlowski

University of Southern California, Los Angeles CA, USA

University of Uppsala, Uppsala, Sweden

Columbia University, New York City NY, USA

Smith College, Northampton MA, USA

Université Laval, Québec, Canada

Monterey Bay Aquarium Research Institute, Moss Landing CA, USA

Dalhousie University, Halifax, Canada

The Natural History Museum, London, UK

University of British Columbia, Vancouver, Canada

University of Geneva, Geneva, Switzerland

Scientific & Technical Advisory Board (STAB)

Linda Amaral Zettler

Claire Gachon

Laure Guillou

Line Le Gall

Laura Wegener Parfrey

Stéphane Audic

Matthew Brown

Micah Dunthorn

Enrique Lara

Frédéric Mahé

Ramon Massana

Alexandros Stamatakis

Woods Hole Marine Biological Laboratory, Woods Hole MA, USA

Scottish Association for Marine Science, Oban, UK

CNRS - Station Biologique de Roscoff, Roscoff, France

Museum National d'Histoire Naturelle, Paris, France

University of British Columbia, Vancouver, Canada

CNRS - Station Biologique de Roscoff, Roscoff, France

Mississippi State University, Mississippi State MS, USA

University of Kaiserslautern, Kaiserslautern, Germany

University of Neuchâtel, Neuchâtel, Switzerland

CIRAD, Montpellier, France

Institut de Ciències del Mar, Barcelona, Spain

Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

ABSTRACT

Universal taxonomic frameworks have been critical tools to structure the fields of botany, zoology, mycology, and bacteriology as well as their large research communities. Animals, plants and fungi have relatively solid, stable morpho-taxonomies built over the last 3 centuries, while bacteria and viruses have been classified for the last 3 decades under coherent molecular taxonomic frameworks. By contrast, no such common language exists yet for protists, even though environmental -omics surveys are showing that microbial eukaryotes make up most of the organismal and genetic complexity of our planet's ecosystems! In fact, the extreme morpho-functional complexity of protists has historically divided the relatively small research community into sub-communities (protozoology vs. phycology, aquatic vs. terrestrial protistology, etc.) speaking different languages. With the current deluge of meta-omics data clearly pushing us into the century of protistology, we urgently need to build up a universal taxonomy bridging the protist -omics age to the centuries-old body of classical knowledge that has effectively linked protist taxa to morphological, physiological, and ecological information. *UniEuk* is a community-based project to address this fundamental challenge and achieve a morpho-genetic reference system for eukaryotic biology, ecology and evolution, including two components. First, a standardized curation process based on the *EukRef* initiative and realized predominantly by PhD students and post-docs will generate phylogenetically-informed reference databases of genetic markers with reference alignments and trees. Second, the generated knowledge will be structured into a universal eukaryotic taxonomic framework, integrating information from relevant genetic markers and classical morphology-based data, and validated by a comprehensive network of taxonomy experts. The system's broad use and long-term preservation will be ensured by a direct implementation into the INSDC genetic data repositories via their EMBL-EBI node.

BACKGROUND

Exploring, describing and understanding the diversity of life on Earth have always been major basic goals of the biological sciences. For various, mostly historical reasons, however, parts of the tree of life have been given much more attention than others. At the macroscopic end of the organismal size spectrum, the diversity of multicellular eukaryotes (plants, animals and fungi) has been extensively described and classified over the last three centuries. Detailed and relatively stable taxonomic frameworks exist for them, essentially based on morphological data, with recent adjustments from genetic evidence. The advent of light- and later electron-microscopy triggered a first major revolution in our understanding of the diversity of life, giving access to the invisible world of viruses, prokaryotes and single-celled eukaryotes. Since then, fascination for the 'real' microbes (prokaryotes and viruses) at the sub-light microscopic end of the organismal size spectrum has captured a lot of the research community's attention. The comprehensive reference genetic databases developed over the last 30 years for prokaryotes – e.g. *SILVA* (www.arb-silva.de), RDP (rdp.cme.msu.edu) or Greengenes (greengenes.lbl.gov) for rRNA genes – take into account data from environmental gene sequencing surveys and are organized into a coherent and powerful molecular taxonomic framework acting as a universal language for the thousands of scientists exploring the biology, ecology, and evolution of bacteria and archaea.

By contrast, attempts to explore the bewildering diversity of protists – which occupy a wide range of organismal sizes from less than one micron to several meters and started diversifying before animals and plants – have remained limited by the relatively small size of the protistan research community. It has been recognized for a long time now that the vast majority of protists are difficult to characterize morphologically and cannot be cultured. Even in the few groups that are extensively studied because of their medical or economic importance, traditional culturing- and microscopy-based approaches only allow access to the tip of the diversity iceberg. The complex physico-chemical and biotic interactions required by most species are typically not reproducible in the laboratory. Despite two centuries of protistology, new lineages of significant ecological and biogeochemical importance are regularly discovered, some of completely novel morphology, others described decades ago but not observed since. For the last few decades, Sanger-DNA sequencing based approaches have been used to explore the diversity of protists from environmental samples, as was previously successfully undertaken with bacteria and archaea. Although efficient at highlighting hidden species within known groups, these 'traditional' molecular ecology approaches remain restricted to limited taxonomic and ecological scales and still overlook most of the protistan genetic diversity.

We are facing today what could be the second most important revolution in our understanding of the diversity of life on Earth since the invention of microscopy. The advent of massively-parallel, next-generation DNA sequencing technologies provides us **for the first time potentially unrestricted access to the global diversity of life in all environments, covering the entire size-range of living communities from viruses to animals**. The rapidly accumulating genetic data from holistic meta-omics approaches are bewildering yet unequivocal – **most of the total biodiversity in the Earth's ecosystems actually belongs to the least studied compartment of life: microbial eukaryotes or protists**. In the world's sunlit open oceans, for instance, as demonstrated by a comprehensive survey of plankton diversity, from viruses to fish larvae, during the circumglobal *Tara* Oceans expeditions (2009-2013), >85% of total eukaryotic rRNA gene diversity corresponds to protists, and this ratio may well hold true for other marine, freshwater, and terrestrial aerobic ecosystems. By extrapolating from estimates of the total number of animal, plant and fungal species on Earth, the total number of eukaryotic species may well be above 100 million. Of note, protists have generally large and complex genomes (some more than humans), and again the global *Tara*-Oceans meta-transcriptomics / -genomics data from the world oceans show that **eukaryotic genes are significantly more diverse and unknown than bacterial, archaeal, or viral genes**. The first large-scale sequencing effort to explore the gene repertoire of cultured protistan strains across large taxonomic scales – the *Marine Microbial Eukaryote Transcriptome Sequencing Project* (MMETSP), supported by the Gordon and Betty Moore Foundation's Marine Microbiology Initiative – further reveals the enormity of our ignorance about protistan genomics, a knowledge gap that will increase dramatically with the advent of environmental single eukaryotic cells transcriptomics / genomics and whole community meta-transcriptomics / -genomics.

Now that the Pandora's box of protist biodiversity has been opened, we are faced with one of the biggest challenges ever in biological sciences and urgently need to give ourselves the means to adequately address, analyze, and understand this staggering diversity. **The coming years represent a critical window to build bridges between the fragile, more than two-centuries-old morpho-physiological knowledge of microbial eukaryotic diversity¹ and the deluge of novel environmental eukaryotic sequence data** generated by next-generation DNA sequencing from single-cells to whole ecosystems. This novel genetic information can only be understood from a functional and ecological point of view if it is linked to the phenotypic (morphological, physiological, ecological) information about the organisms in question. In order to understand the critical role of unicellular eukaryotes in ecosystem functioning, as well as their tremendous yet unexploited potential for bioresources, biotechnology, and biomonitoring, we need to set up a strategy to rapidly construct an efficient **universal morpho-genetic reference system for protists**. To this end, existing and future reference databases of protistan genes must be integrated into **a common universal taxonomic framework**. Only then will we be able to efficiently assign organismal identity and phenotypic information to environmental sequence data. This will **provide a common language for the fast-growing protistan scientific community, while preserving two centuries of classical protistology that have effectively linked taxon names to morphological, physiological, behavioral, and/or ecological information**.

None of the currently available databases and taxonomic frameworks can adequately address the challenges we are facing with protists. Databases in the International Nucleotide Sequence Database Collaboration (INSDC: NCBI GenBank, DDBJ, and EMBL-EBI/ENA) for instance lack a reliable taxonomic framework for unicellular eukaryotes and contain many taxonomic assignment errors. This is especially problematic in the case of environmental sequences that become biologically/ecologically relevant only when correctly assigned to the organism they come from. Other databases focusing on a single gene may better integrate taxonomic information, but here the risk is that each has its own taxonomic framework, making it impossible to link them into a common morpho-genetic reference system. In fact, various taxonomic frameworks do exist for protists, the most recent and ambitious being that proposed in 2012 by the International Society of Protistologists², based mostly on merging morphological and phylogenetic data. The co-existence of distinct competing taxonomic frameworks is detrimental to the task at hand, which needs a universally accepted consensus classification, and the frameworks are currently not linked to genetic information, thus missing a vast part of protistan biodiversity not yet described morphologically, but known from environmental DNA surveys (essentially 18S rRNA gene surveys).

¹ Soon to be lost in the mass of largely non-translated, non-digitalized protistan knowledge printed in German, Russian, French, etc. in journals and books of difficult access, or through retirement of rare and non-renewed taxonomy experts.

² Adl et al. (2012) The Revised Classification of Eukaryotes. *J. Eukaryot. Microbiol.* 59:429-493.

As a result, **various community efforts have emerged to improve the situation by developing databases and tools at the interface between genes and morphological knowledge**. Starting in 2009, the International Census of Marine Microbes developed the hypervariable V9 region of the eukaryotic rRNA gene as the choice barcode for NGS protistan surveys, used later for a global ocean assessment of planktonic eukaryotes in the Tara Oceans expeditions. A 'Protist Working Group - ProWG', including 33 taxonomy experts and covering most major protistan lineages, was created in 2011 as part of the Consortium for the Barcoding Of Life (CBOL) and proposed a barcoding strategy that accommodates the biological and evolutionary complexity of protists. This strategy consists of the use of a universal eukaryotic pre-barcode (the hypervariable V4 region of the 18S rRNA gene) allowing a first-order taxonomic assignment into a universal framework, and orientation towards lineage-specific, species-level barcodes³. On the other hand, the Protistan Ribosomal Reference (PR²) database⁴ focuses on 18S rRNA genes and was built in 2013 within a taxonomic framework curated by the *BioMaRks* European consortium of experts (<http://biomarks.eu>). PR² was a promising start, but needs regular updating and curation and in-depth revision in several taxonomic groups. In addition to bacteria and archaea, *SILVA* includes eukaryotes, but also requires a more reliable taxonomic framework for protists, which led to the recent creation of a eukaryotic taxonomy working group (<http://www.arb-silva.de/projects/eukaryotic-taxonomy/>).

To our knowledge, at least three mostly PR²-inspired projects of curated 18S rRNA gene databases were proposed in the last year⁵, all putting forward the idea of a community of contributors (mostly PhD students and postdocs) curating protistan gene sequences to build reference gene databases, the taxonomy of which should be validated by experts. In addition to these 18S rRNA gene database projects, some research groups are developing additional morpho-genetic frameworks based on other nuclear or organellar gene markers, such as the PhytoREF database of eukaryotic plastidial 16S rRNA gene reference sequences (~6,500 taxonomically curated sequences covering all known photosynthetic eukaryotes⁶). Such initiatives would need to be integrated into a single universal and coherent taxonomic framework to maximize cross-information and fertilization. Each of these projects represents a clear improvement and should be strongly supported. However, **we believe that coordination of these efforts into a single, universal morpho-genetic reference system would be hugely beneficial to the whole scientific community and will save a tremendous amount of time and energy**. The worst possible scenario is where we are today, with multiple competing reference databases, with slightly different goals in mind and distinct taxonomy. What the scientific community needs at this stage is a shared and universally supported morpho-genetic reference framework for protists that will allow us to confidently enter the new era of 'meta-protistomics'.

UNIEUK GOALS

We propose to develop, **within the next 3 years, a morpho-genetic reference system integrating an adaptive universal taxonomic framework for eukaryotes with reference gene sequence databases** (Fig. 1). It will be organized as a **community-based annotation effort of various reference gene databases (bottom-up curation** based on the *EukRef* concepts and bioinformatics pipeline (<http://eukref.org/>)), and integrated into a **taxonomic framework validated by a comprehensive network of taxonomy experts (top-down validation)**. This system will allow continuous growth and updating over time and should in the long-term be self-sustainable. The main objective of the project is the creation of long-lasting, interconnected and curated online public reference databases for eukaryote biology, ecology and evolution (phylogeny, taxonomy, barcoding, environmental surveys, biomonitoring, etc.) all fitting into a common universal taxonomic framework. **The system will include two major components, which will be fully integrated, but will represent separate, stand-alone end-products that will be available to the scientific community and other end-users:**

³ Pawlowski *et al.* (2012) CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biology* 10:e1001419.

⁴ Guillou *et al.* (2013) The Protist Ribosomal Reference database (PR²): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 41:D597-D604.

⁵ From Philippe Deschamps and Purificación López-García (Paris), Laura Wegener-Parfrey, Javier Del Campo, Matt Brown, and Colombar de Vargas (Vancouver), and Cédric Berney, Stéphane Audic and Colombar de Vargas (Roscoff).

⁶ Decelle *et al.* PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources*, in press.

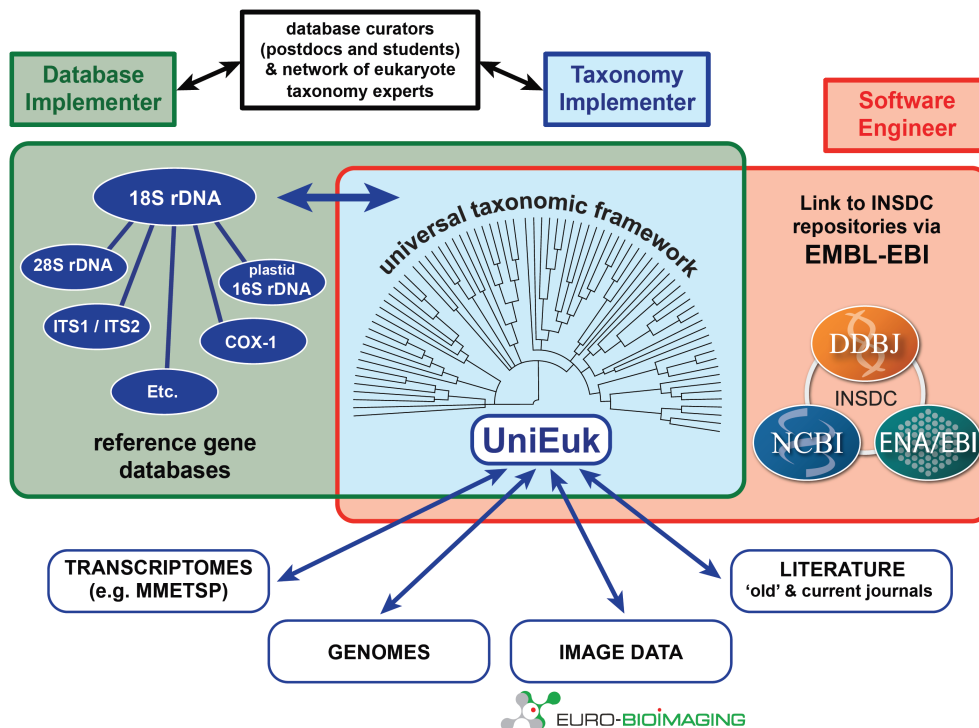


Figure 1.
UniEuk implementation scheme

(i) An organism-based, **universal eukaryotic taxonomic framework** integrating classical, morphology-based taxonomic knowledge and phylogenetic information from the 18S rRNA gene and other genetic markers (blue rectangle in Fig.1). The framework needs to be consensual, flexible, adaptive, and efficient to fit the needs of the protistology community and beyond, and make it possible to co-opt existing taxonomic systems for multicellular eukaryotes (animals, plants, and fungi). It should be agreed upon and adopted by the main actors in the field, such as protistological and phycological societies and their corresponding journals. With support from the INSDC genetic data repositories and current eukaryotic reference gene databases such as PR² and SILVA⁷, this framework will become the prevalent taxonomic system for eukaryotes. The consolidated taxonomic framework will be propagated back to INSDC databases through two routes. First, within the timeframe of the project, the EMBL-EBI node will integrate the taxonomy itself and the sequence classifications under that taxonomy into their permanent record, as a searchable and navigable classification system across protist sequence records. This first phase has a guaranteed output and will leave a permanent mark on the world's data, being preserved at EMBL-EBI beyond the duration of the project. Second, drawing on the direct experience of serving data in the new framework, we will work to establish flow of the new framework into the established NCBI Taxonomy in the longer-term. The system will have a fixed structure and follow a set of rules clearly defined in advance by the community in light of the needs of end-users, allowing its long-term sustainability and integration into the INSDC databases. However, it will also be fully flexible and adaptive in its arborescence, allowing regular updates whenever new data become available or as the community's views on the taxonomy of certain groups change over time.

(ii) **Curated reference gene sequence databases** that will contain phenotypic and environmental contextual data encoded using existing metadata standards (cf. <http://gensc.org>) (green rectangle in Fig.1). Individual reference databases will be organised under the universal eukaryotic taxonomic framework, and used to refine it iteratively. Because of the historical importance and primary use of 18S rRNA gene sequences in assessment of environmental diversity, providing an updated 18S rRNA gene reference database is the first goal of the project, and can be seen as a case study for the overall approach. The 18S rRNA gene reference database will be the core, central component of the system, as recommended by the CBOL protist initiative. However, existing or future reference databases for other genes (e.g. 28S rRNA and nuclear, chloroplastal or mitochondrial single-copy genes) that can be used for eukaryote phylogeny, environmental monitoring, group-specific barcoding, etc. will be integrated into the system to further improve the universal taxonomic framework and ensure the long-term unity of the whole system.

⁷ But also RDP, Open Tree of Life, Algaebase, CBOL, EOL, the Global Taxonomy Initiative, WoRMS, etc.

UNIEUK ORGANIZATION

The project will be led by a **Steering Committee (SC)** under the supervision of an **Advisory Council (AC)**, while practical input will be provided by a **Scientific and Technical Advisory Board (STAB)** both at the outset and during the course of the project. **Three implementers** will ignite and maintain the energy and cohesiveness of this international effort over a period of 3 years: a **Taxonomy Coordinator** (full time), a **Database Coordinator** (full time), and a **Software Engineer** (years 2 and 3) (Fig. 2).

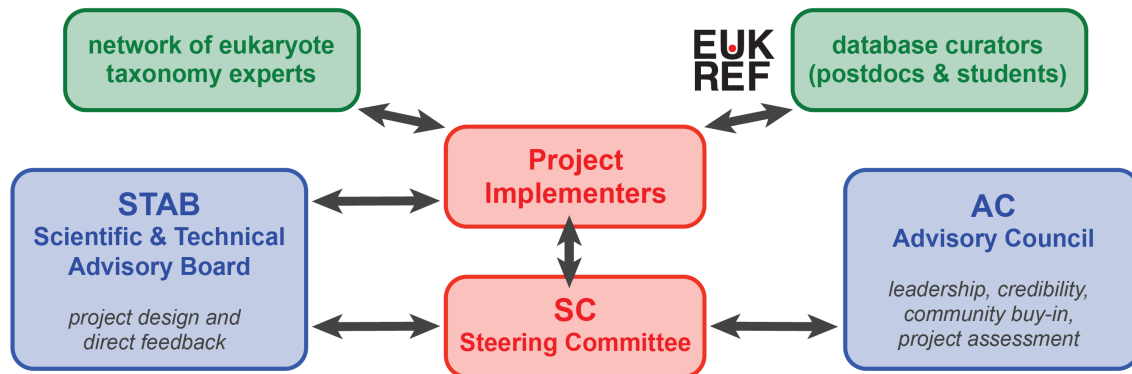


Figure 2. UniEuk organizational scheme

The **Steering Committee (SC)** is the decision-making body of the project. It includes a Director and an co-Director acting as the main contact points for communication purposes, and is composed of members with complementary expertise: coordinators of the existing initiatives in eukaryote taxonomy and reference gene sequence databases, experts in protist taxonomy, ecology, and evolution, and representative of the INSDC gene sequence repositories. In case of conflict it will have final say on decisions by majority vote.

Roles of the SC:

- Lay out the project's precise goals and timeline, with input from the STAB and the AC, especially in the initial phase of the project;
- Oversee the project's progress and suggest ways to implement recommendations from the AC;
- Liaise between the implementers, the software engineer, and the STAB and AC;
- Be responsible for the communication strategy for the project;
- Raise funds from national and private research funding agencies and international protistology societies to cover the costs of additional database curation workshops or small projects to generate reference gene sequences for key taxonomic lineages underrepresented in the databases.

In addition to substantial contributions in kind by the participating institutions (mainly CNRS in France, Max Plank Institute and Jakobs University in Germany, and EMBL-EBI in the UK), the International Society of Protistologists ISOP (<http://protozoa.uga.edu/>) has already committed \$30,000 to *UniEuk* to demonstrate support for the project. Other potential sources of funding could include:

- EU CSA (http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/annexes/h2020-wp1415-annex-d-csa_en.pdf)
- US NSF programs Dimensions of Biodiversity or Systematics and Biodiversity Science cluster
- Human Frontier Science Program (<http://www.hfsp.org>).
- SCOPE program at the Simons Foundation (<https://www.simonsfoundation.org/life-sciences/simons-collaboration-on-ocean-processes-and-ecology/>)

The **Scientific and Technical Advisory Board (STAB)** is composed of experts in the fields of genetic reference database management and eukaryote taxonomy and barcoding.

Roles of the STAB:

- During the project's initial phase, provide practical suggestions about the project's implementation;
- Over the course of the project, provide rapid, direct feedback whenever questions / problems arise.

The **Advisory Council (AC)** is composed of key leaders in the fields of protistology and representatives of the major stakeholders of the project (protistological and phycological societies, INSDC gene sequence repositories, the GBMF project MMETSP, etc.).

Roles of the AC:

- Provide recommendations to ensure that the project's outcomes match the needs of the scientific community and of the stakeholders;
- Ensure the scientific leadership and credibility of the project, as well as community buy-in;
- Convene annually at international protist conferences to assess the project's progress;
- Liaise with protistological and phycological societies, culture collections, etc. to promote the project and the universal taxonomic framework across the scientific community.

The **Taxonomy Coordinator** (Station Biologique de Roscoff, France) will:

- (i) follow the bottom-up curation of the databases and address taxonomic questions emerging from the curators based on guidelines from the *UniEuk* boards;
- (ii) integrate morphological evidence compiled from the literature and output from *EukRef* (18S rRNA gene) and other genetic reference databases into the *UniEuk* taxonomic framework and communicate with the network of taxonomy experts for validation of all suggestions;
- (iii) use feedback from the taxonomy experts to guide future curation efforts and iteratively optimize the framework across all eukaryotes;
- (iv) ensure that satellite databases (other marker genes) are implemented into the system without taxonomic conflict;
- (v) participate in the production of outreach material to help the SC and AC to promote *UniEuk* to the scientific community via publications in peer-reviewed journals, presentations at international meetings, etc.; and
- (vi) liaise with the Database Coordinator to facilitate integration of the reference gene databases into the common taxonomic framework.

The **Database Coordinator** (Max Planck Institute & Jacobs University, Bremen, Germany) will:

- (i) implement the database model to handle sequence data and taxonomic information (see below - proposed implementation);
- (ii) develop scripts for annotation and data management including the status of a Taxon;
- (iii) supervise the creation of the web portal interface for data visualization and annotation in close cooperation with the Taxonomy Coordinator;
- (iv) participate in the production of outreach material to help the SC and AC to promote the reference gene sequence databases to the scientific community via publications in peer-reviewed journals, presentations at international meetings, etc;
- (v) liaise with the Taxonomy Coordinator to help improve the common taxonomic framework via feedback from the reference gene databases; and
- (vi) liaise with the Software Engineer to facilitate propagation of the system into EMBL-EBI data structures.

The **Software Engineer** (EMBL-EBI, Cambridge UK) will:

- (i) support discovery, curation and presentation of protist sequence data within INSDC that feed the project's taxonomy construction workflows;
- (ii) establish, in collaboration with the SC and Coordinators, a format for the efficient representation of the taxonomic framework and the classification of sequence data in the framework;
- (iii) establish a workflow to allow routine import of this representation into EMBL-EBI data structures;
- (iv) establish data indexing workflows to allow web and programmatic discovery and browsing in the context of the framework of sequence records classified therein;
- (v) construct and deploy a browser and programmatic taxonomic entry, navigation and visualisation point for protist data classified in the framework; and
- (vi) work across INSDC to establish the *UniEuk* framework as a primary feed into the NCBI Taxonomy.

Both coordinators and the software engineer will report to members of the SC, and the SC will communicate progress reports to all members of the AC and the STAB. The SC will meet with the STAB and the AC at the start of the project to clearly lay out its objectives and the main expected outcomes in light of the needs of the scientific community and the stakeholders. The AC will then convene yearly with the SC and implementers (and optionally the STAB and lead taxonomy experts) to assess the project's progress, deal with possible issues, and discuss further actions. A final joint SC / AC / STAB meeting will take place at the end of the project (see Annex 1: *UniEuk* TimeLine).

UNIEUK IMPLEMENTATION

Our primary vision of how the project could be implemented is given in the following paragraphs as a strawman. However, from the inception of the project and throughout its duration constant input will be sought to optimize and refine the structure of the overall process, from the STAB and AC, and the whole community involved. This input began in July 2015 during the first *EukRef* workshop in Vancouver organized by L. Wegener Parfrey, J. del Campo, M. Brown, and C. de Vargas (*EukRef*, the 18S rRNA gene collaborative annotation initiative, <http://eukref.org/workshops/vancouver-workshop/>). By that time all members of the STAB and the AC had confirmed their interest in participating in *UniEuk*. Discussion to refine *UniEuk*'s implementation plan was held between members of the SC, AC, and STAB attending the workshop. Outreach to and feedback from the wider community began at the ECOP/ISOP joint meeting in Sevilla (September 2015, <http://www.viiecop.com/>), where the *UniEuk* and *EukRef* projects were jointly presented to the scientific community during a special mini-symposium, followed by the first *UniEuk* community-meeting on Sept. 9; members of the STAB and AC, and lead taxonomy experts were invited to attend this half-day meeting to gather formally with members of the SC. General feedback from the community on the proposed implementation plan (Figure 3) and timeline (Annex 1) was extremely positive and constructive for both the *EukRef* and *UniEuk* projects.

Figure 3. *UniEuk* implementation workflow

The Universal Eukaryotic Taxonomic Framework is built through bottom-up / top-down, standardized, semi-automated curation processes involving the whole scientific community, and implemented at EMBL-EBI with regular updates.

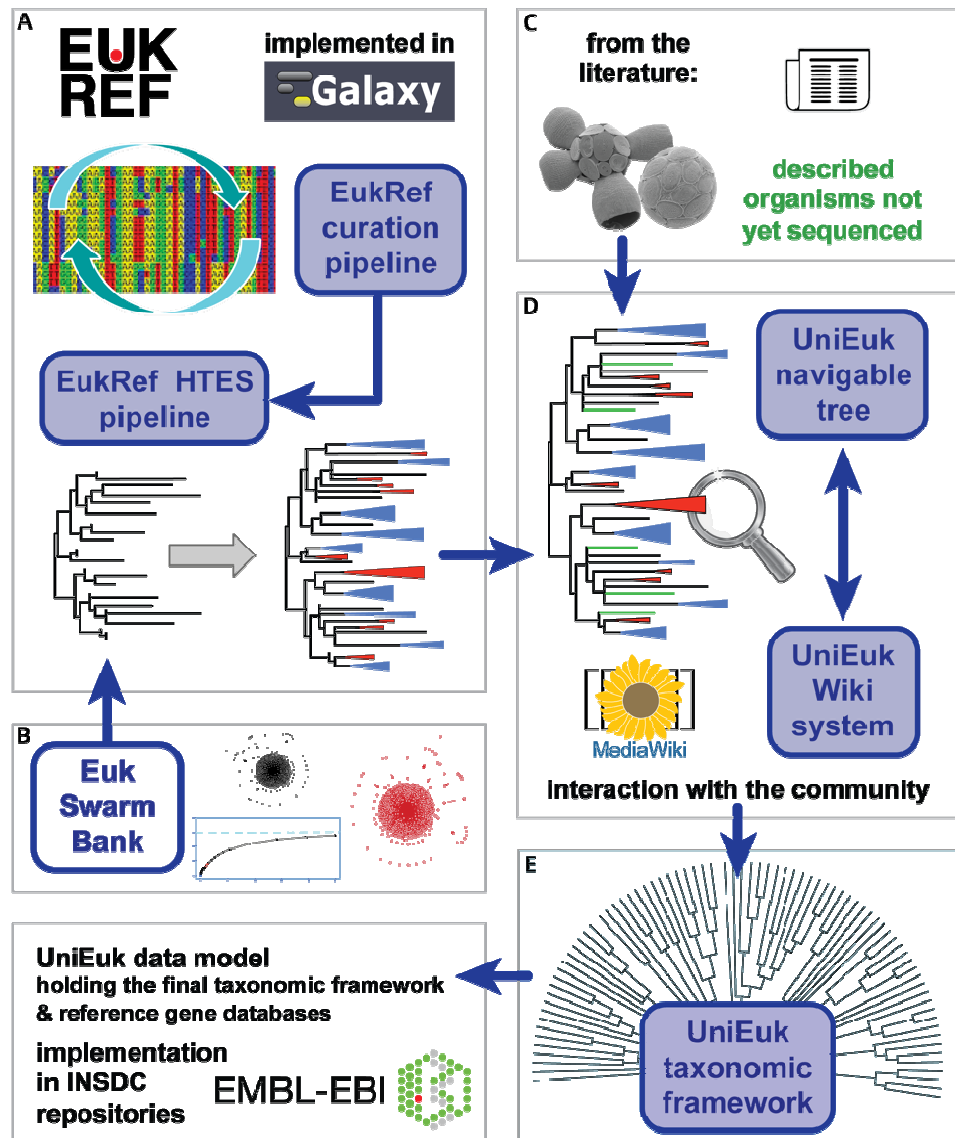
A. The *EukRef* curation & HTES (High-Throughput Env. Sequences) pipelines (<http://eukref.org/pipeline/>), implemented in Galaxy, allow standardized curation of reference 18S rRNA gene alignments and trees by the research community, on a group-by-group basis. Red and blue branches represent Sanger and HTES environmental sequence data, respectively.

B. The Euk_Swarm Bank ingests all novel V4 rRNA gene data, swarms it, and sorts out new eukaryotic diversity to be inserted into EukRef Trees.

C. Information from purely morphological data is also inserted into the UniEuk Guide Tree (green branches).

D. The UniEuk Guide Tree is navigable and linked to a wiki system (each node is a clickable entity), allowing communication between the community and the project implementers.

E. The UniEuk Taxonomic Framework is connected to various marker genes tables, and implemented at EMBL-EBI with regular versioning.



(i) *EukRef* marker genes curation pipelines and *UniEuk* Swarm Bank

The 18S rRNA gene reference database will be the main source of genetic information to inform the organisms-centered *UniEuk* taxonomic framework (taxoframe). Other genetic markers will be used to inform the taxoframe wherever the resolving power of the 18S rRNA gene is reaching its limits. Curation of the 18S rRNA gene database started during the first *EukRef* workshop at UBC (Vancouver, Canada) in July 2015. The *EukRef* curation pipeline demonstrates the power of the proposed standardized bottom-up semi-automated curation approach. The principle of a network of curators ensures that the people who have the most input into reference databases – essentially PhD students and postdoctoral researchers – are also the ones who benefit most from them for their own research. Follow-up *EukRef* reference gene database curation workshops will be organized during the first year of *UniEuk* (see Time Line, Annex 1), to cover lineages not yet included in the 18S rDNA database, or for other marker genes. The *EukRef* curation pipelines will also be made available via *Galaxy* so that they can be used remotely by curators anywhere and can start to be applied to other marker genes to curate additional genetic reference databases with the same *EukRef* standards.

In practical terms, reference sequences from all major eukaryotic groups are used to build starting alignments at different taxonomic levels, allowing extraction of novel, clade-specific gene sequences from the INSDC databases. After removal of chimeras and sequences shorter than a decided threshold, the novel sequences are aligned to their reference counterparts, phylogenetic trees are constructed to inform their taxonomy, and the new alignments are used to repeat the searches. The process repeats iteratively until no more new INSDC sequences can be added to the corresponding alignment. This iterative *EukRef* process provides output material in the form of reference alignment and phylogenetic trees. Throughout this process curators are encouraged to use their knowledge of the group to better inform curation, for example by using information from multigene phylogenies and morphology to optimise interpretation of the phylogenies. To ensure that information from environmental surveys is implemented into the taxonomic system from the start, simple and practical naming guidelines were agreed upon in Vancouver (by attending members of the SC, STAB, AC and lead taxonomy experts). These rules are especially important to annotate the increasingly abundant environmental 18S rRNA gene sequences belonging to genetically known but morphologically undescribed lineages. This system will be refined over the course of the *UniEuk* project.

The *EukRef* process includes a specific pipeline for the taxonomic annotation of data from high-throughput environmental sequencing (HTES) studies, which maps them onto the reference phylogenetic trees that are the main output of the curation pipeline. Data from the *Tara* Oceans project and other HTES efforts agreeing to contribute their data to the *EukRef* / *UniEuk* effort will be distributed on a group-specific basis to curators responsible for each eukaryotic lineage. The growing 18S rRNA gene reference database will be used to annotate and analyse these data and should lead to many publications highlighting novel diversity within clades and/or distribution of taxa across environments. In addition, when the sequence data generated by such studies is phylogenetically informative, it will be useful to integrate them into the *UniEuk* taxonomic framework together with the information from the *EukRef* reference phylogenetic trees. However, such data can be highly redundant, therefore it would be impractical and unwise to try and blindly incorporate them all into the *UniEuk* framework. Rather, we propose the creation of a "Euk_Swarms Bank" relying on the powerful, recently published Swarm clustering method to generate OTUs from next-generation sequence data⁸. For any targeted gene fragment (such as the V4 or V9 regions of the 18S rRNA gene), sequence data from HTES studies will be rationalized by sorting them into Swarms, while retaining all useful information about their distribution and abundance in the original datasets. Only non-redundant, phylogenetically informative Swarms will then be selected to be integrated into the *UniEuk* taxoframe.

(ii) *UniEuk* taxonomic framework and wiki-guide tree for interaction with the community

The universal eukaryotic taxonomic framework (*UniEuk* taxoframe) should represent a community-wide consensus and will be validated by a comprehensive network of eukaryote taxonomy experts, on a group-by-group basis. It should be the result of a fundamentally collaborative effort involving the whole protistology community, which is still of manageable size and already has strong ties thanks to regular international meetings and the existence of large-scale collaborative projects such as ProWG, *BioMarks*, *Tara-Oceans*, *LifeWatch*, GBif, MMETSP, etc. To ensure consensual agreement on the framework, no single eukaryotic lineage should be the responsibility of a single taxonomic expert, and the taxonomy of any lineage

⁸ Mahé et al. (2014) Swarm: robust and fast clustering method for amplicon-based studies. *Peer J.* 2:e593.

should be subjected to approval by all experts, with disagreements solved on a majority rule basis. The taxonomic framework should match known or hypothesized evolutionary relationships as closely as possible, so that it can be used by the whole scientific community and in textbooks. Its main characteristics should nevertheless remain its flexibility, updatability, and practicality, given that one of its main goals is to be a tool for sequence data annotation and a unique link between past and future protistology. Taxonomy experts who already participated in the system proposed by the International Society of Protistologists⁹ and/or were part of the ProWG, PR²/BioMarkS, the SILVA Eukaryotic Taxonomy Working Group, or Tara-Oceans and MMETSP projects have already been contacted, providing the seed to quickly extend the network further throughout the whole protistology community, via mailing lists, special sessions at international protistology and phycology meetings, etc. This will also help to build community interest, buy-in, and participation.

With the help of the network of taxonomy experts, work on the *UniEuk* taxoframe will start immediately at the beginning of the project in terms of stabilizing the higher-level classification of eukaryotes, integrating information for lineages whose 18S rRNA gene sequences have already been curated with the *EukRef* pipelines, and adding information from the literature for described but not yet sequenced taxa. This work will happen while the final *UniEuk* data model is being created (see below); in the meantime a straightforward but efficient feedback / communication / validation system will be made available to all curators and experts. Initial interaction with the community can occur via group-specific mailing lists and document-sharing tools to achieve a starting version of the taxonomic framework. Within six months of the start of the project, this primary framework will be presented online to the community so that more precise and direct feedback from the whole community becomes possible. For this we propose to use a wiki system linked to a navigable tree representing the taxonomic framework, which will be available upon registration to the project via the *UniEuk* web portal. This guide tree will allow visualization of the framework from the root to the terminal nodes. Each internal and terminal node will later correspond to an entry in the Taxon Table of the *UniEuk* data model (see below), and will be attributed a unique *UniEuk* identifier. Each node in the guide tree will correspond to a wiki-page where curators and experts will be able to provide information, make suggestions of additions and changes, propose solutions when questions or conflicts arise, etc. The Taxonomy Coordinator will be responsible for moderating the decision-making process and integrating the suggestions made by taxonomic experts into the unified taxonomy. Creation of the *UniEuk* wiki-tree will therefore be a high priority to facilitate and catalyze interaction between the community and Taxonomy Coordinator immediately after the starting version of the framework is ready.

(iii) *UniEuk* data model and web portal

A tentative data model for the *UniEuk* project has been proposed by the Steering Committee. It is based on the following agreed concepts: (i) the model is designed to hold the final *UniEuk* taxonomy only; (ii) changes made to the unified taxonomy's taxa will be tracked; and (iii) the decision process and conflict resolution will be managed by the Taxonomy Coordinator. These requirements allow for a straightforward data model, focusing on the taxa of the *UniEuk* taxoframe and avoiding advanced concepts like authorization ("who changed what"), a full audit trail implementation ("who changed what when"), and working copies for each taxonomic expert used to create proposals for updates/corrections to the taxonomy. Separate tools will be used (i.e. the *UniEuk* WikiTree mentioned above) to support the curation of the taxonomy, the decision-making process, and conflict resolution within the *UniEuk* project. The data model will serve to hold the unified taxonomy, and all the associated information. It will also be utilized to generate different outputs of the *UniEuk* project. These products will be made available in regular release intervals. A pragmatic approach with short cycles between users and developers will be followed for the first roll-out of the database and web interface, before completeness of features. This should allow implementation of a first usable prototype within 12 months of the start of the project. Implementation of the *UniEuk* taxonomic framework in the INSDC repositories via EMBL-EBI will then take place during year two, with a proposed subsequent versioning every six months. The proposed data model (Fig. 4) is described below. A prototypical implementation of this data model based on a relational schema (implemented in MySQL) is already used by the SILVA project to assign unique numerical identifiers and to track changes to taxa in the SILVA taxonomy (green tables in Fig. 4). It does not currently include Traits, Metadata, and Synonyms (orange tables in Fig. 4). If needed, the data model can be further simplified: fundamentally necessary tables are the **Taxon Table** and the **Sequence Table**; all other information could be implemented as optional fields.

⁹ Adl et al. (2012) The Revised Classification of Eukaryotes. *J. Eukaryot. Microbiol.* 59:429-493.

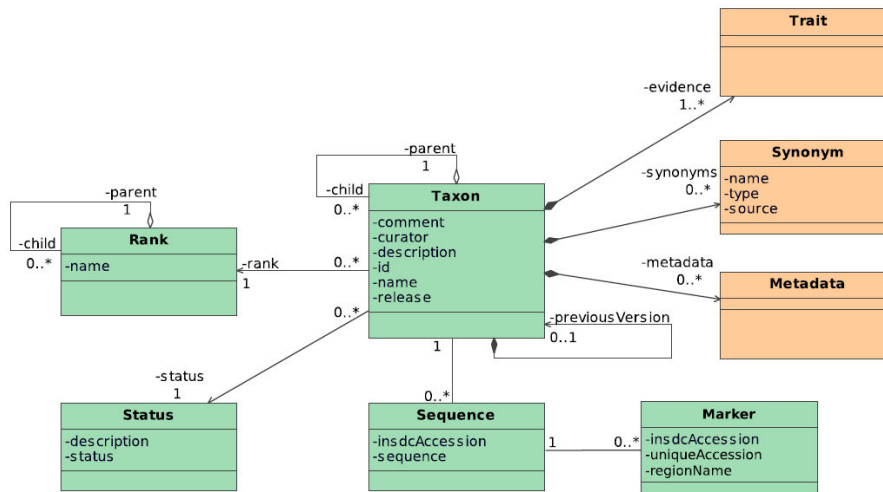


Figure 4. Tentative data model of the *UniEuk* taxogenetic database

The data model is centred around the **Taxon Table**, which holds the key taxonomic data provided by the coordinators, curators, and experts. A Taxon here refers not only to species, but also to any taxonomic rank or phylogenetic depth both below and above species, i.e. strain ATCC 50883, *Amphitrema wrightianum*, Labyrinthulea, and Stramenopiles are all taxa and will be in the central **Taxon Table**. Each Taxon will have a numerical ID that can be used to uniquely identify it within the *UniEuk* taxonomy; IDs will never be re-used, even if a Taxon is not used anymore. Whenever a Taxon is changed, a link to its previous version can be created, recording its history. The links between a Taxon's parent node and its (multiple) child taxa will also be established via the ID. In addition to the Taxon's name and ID, the **Taxon Table** can also hold optional fields such as the name of the Taxon's curator, a comment/description, in which release it was added to the taxonomy, or the taxonomic rank it is generally given. A **Status Table** can be used to describe the status of a Taxon (published, obsolete, deleted, merged, split, renamed, etc.). This will ensure that no Taxon name is ever lost, and changes can be easily tracked. This would prove particularly useful if mappings between older and newer versions of the *UniEuk* taxonomy are to be created. The evidence that supports a Taxon and its positioning in the taxonomy can be represented by the **Trait Table**, which may include in particular phenotypical, physiological, and ecological information, as well as information about the most appropriate phylogenetic marker for the Taxon. Taxa may have different names for both historical and biological reasons. The German Collection of Microorganisms and Cell Cultures (DSMZ), for example, curates a list of basonyms, homo- and heterotypic synonyms of microorganisms. Optionally, this information can be represented by a **Synonym Table**. Any kind of additional data that the curators want to assign to a Taxon can be included in a **Metadata Table**.

Because phylogenetic information is central to the *UniEuk* system, all genetic information should be implemented in a separate **Sequence Table** that will provide the necessary link between genetic data and the Taxon it belongs to. In its simplest form, the **Sequence Table** must contain a sequence's unique identifier (typically, the INSDC accession number for sequences that are held in the INSDC repositories) and the ID from the **Taxon Table** to link the sequence to the Taxon it belongs to. In addition, the table can contain a name or description for the sequence, or the actual sequence. The fundamental principle of *UniEuk* is that different phylogenetic markers (whether a whole gene such as 28S rRNA gene or alpha-Tubulin, or only part of a gene such as the V4 or V9 regions of 18S rRNA gene) will be unified under the same taxonomic framework. This can be achieved by the use of distinct **Marker Tables** (one for each possible phylogenetic marker) that will list for each marker the identifiers from the **Sequence Table** of all relevant sequences containing the marker. Whenever the marker is only part of the total sequence (e.g. for genomic contigs containing many genes, or for markers that are only part of a gene), the sequence region information using "start" and "stop" positions. Alternatively, the information contained in a **Marker Table** could be included in the **Sequence Table** as optional fields. However, it may be more convenient to keep the **Marker Tables** as separate entities, given that different contributors may work on different markers at different times, and each **Marker Table** would represent a useful, stand-alone *UniEuk* output for end-users.

A web portal will be created to present the project and the morpho-genetic reference system. It will list all collaborators and their expertise, and offer a user-friendly interface to contact the project's coordinators. The portal will present the universal taxonomic framework for eukaryotes with regular updates, and provide tools allowing feedback from the community and taxonomy experts via a navigable tree and user-friendly wiki system. It will also provide links to the various gene reference databases and their contextual metadata, and information about which gene(s) is/are the most appropriate genetic marker(s) at all taxonomic levels in the framework. Online services will allow curators to upload new annotated reference sequences and flag putative misannotations in each database. In relation to the 18S rRNA gene reference database in particular, reference alignments and phylogenetic trees will be made available for all curated taxonomic lineages. Over the course of the project, special sessions will be organized at international protistology and phycology meetings to promote *EukRef* and *UniEuk*, and present major advances (taxonomic framework, reference databases, and publications deriving directly from them). These sessions will be open to all contributors and participants at the meetings, to encourage new postdocs and PhD students to join the community effort.

(iv) *UniEuk*-related activities

During the course of the project, the SC will forge links to other initiatives relevant to, or which could benefit from, the *UniEuk* taxonomic framework and reference gene databases. The SC will also endeavor to raise funds from national and private research funding agencies and protistology / phycology societies to cover the costs of *UniEuk*-related activities, such as taxonomically targeted sampling projects to generate reference gene sequences for key lineages that are underrepresented in the databases, or projects aiming at uncovering new diversity in under-sampled environments. For practical reasons, we do not envisage *UniEuk* to systematically include imaging data for all taxa. However, existing imaging data can be linked to entries in the taxonomic framework as metadata. In addition the SC, STAB and AC will discuss ways to integrate imaging data into the system during the course of the project, notably by liaising with existing international collaborative bioimaging initiatives (e.g. Euro-BiolImaging, <http://www.eurobioimaging.eu/>)

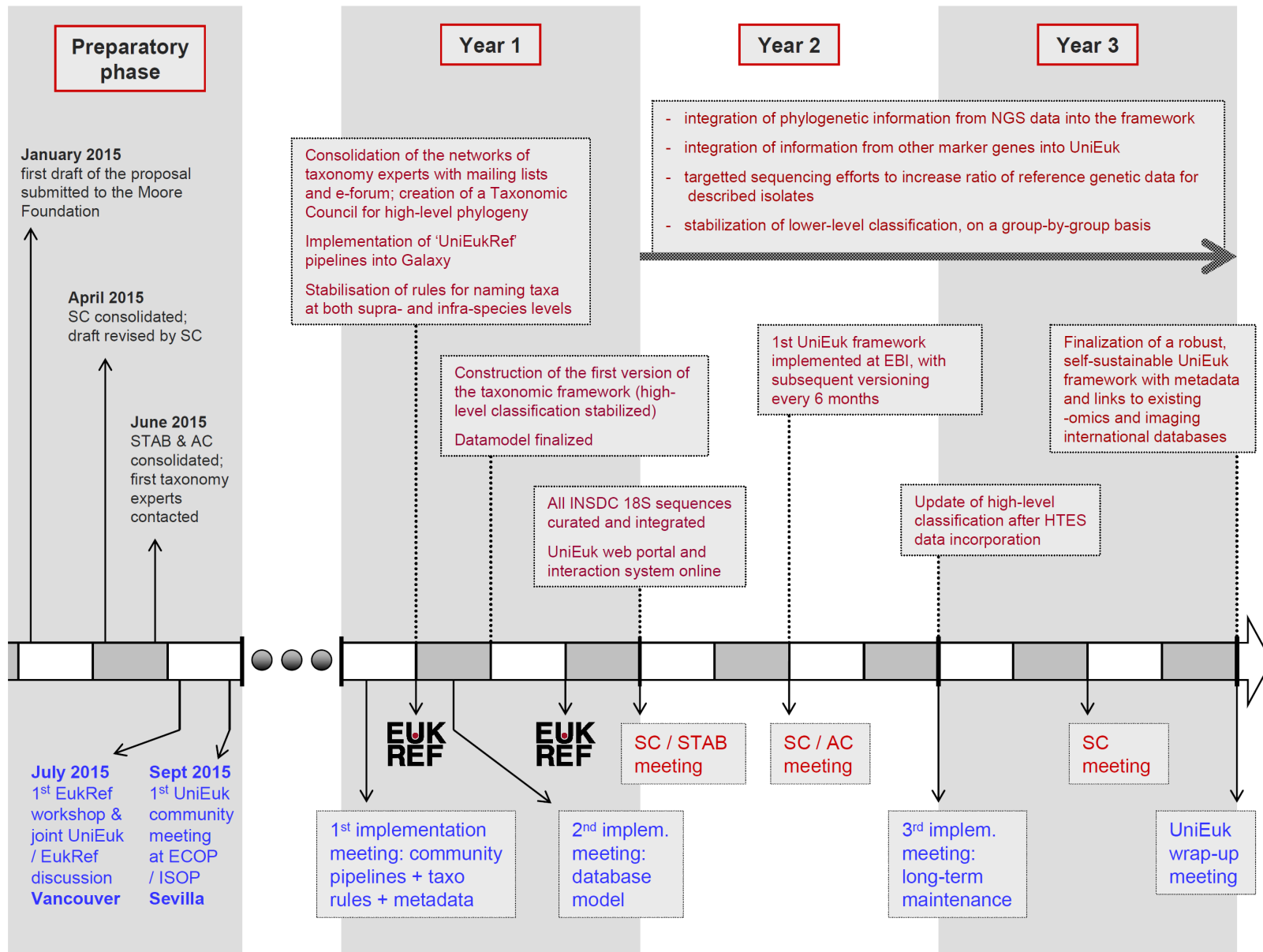
UNIEUK OUTCOMES

The project will foster major scientific outcomes and help address critical research questions:

- *UniEuk* will provide the first unified framework for eukaryotic biology, evolution and ecology, that will be permanently incorporated into the INSDC gene sequence repositories because of the direct involvement of EMBL-EBI; this unified framework will significantly increase cross-fertilization of results and ideas from different research fields ranging from alpha taxonomy to ecology or metagenomics / metatranscriptomics.
- *EukRef* outputs (reference alignments and trees) and curation pipelines will represent highly valuable community resources and set quality standards for future genetic marker reference databases.
- *UniEuk* will improve our vision of the tree of eukaryotic life, by identifying and naming novel and/or deep protistan lineages that will help resolving the origin and macro-evolutionary history of eukaryotes.
- as a reference matrix in analyses of NGS marker gene surveys and metabarcoding datasets (e.g. ICoMM, MIRADA-LTERS, *Tara* Oceans, Malaspina, OSD, etc.), *UniEuk* will allow accurate taxonomic identification of the main ecological players among protists in all environments. These most abundant eukaryotic cells in natural systems should become the primary targets for future single-cell -omics / phenotypic studies (but are typically not among the current eukaryotic model taxa).
- with optimised taxonomic annotation provided by *UniEuk*, network analyses of DNA metabarcoding datasets will allow the discovery of new eukaryote-eukaryote interactions and symbioses in all environments and highlight lineages of particular interest from the point of view of emerging threats to economically critical human activities (agriculture, aquaculture, etc.).
- *UniEuk* will provide a centralised source of information for future projects on eukaryote diversity surveys, assessment of the effects of human activities and climate change on biodiversity, and biomonitoring, especially for newly discovered lineages of medical, ecological, and/or economical significance.
- Finally, the process of developing *UniEuk* will generate a list of key scientific questions for further investigation by the research community.

UniEuk - project proposal

Annex 1: proposed Time Line of the UniEuk project



Annex 2: preliminary list of lead taxonomy experts who have accepted to join the effort; they will help organize sub-groups of experts to cover the whole diversity of protists, and will act as main contact points for the Taxonomy Coordinator during the project.

Sina Adl	University of Saskatchewan, Saskatoon, Canada
David Bass	The Natural History Museum, London, UK
Jens Boenigk	University of Duisburg-Essen, Essen, Germany
Ivan Cepicka	Charles University, Prague, Czech Republic
Jackie Collier	Stony Brook University, New York City NY, USA
Olivier De Clerck	Ghent University, Ghent, Belgium
Micah Dunthorn	University of Kaiserslautern, Kaiserslautern, Germany
Bente Edvardsen	University of Oslo, Oslo, Norway
Laure Guillou	CNRS - Station Biologique de Roscoff, Roscoff, France
Mona Hoppenrath	Forschungsinstitut Senckenberg, Wilhelmshaven, Germany
Sergey Karpov	St. Petersburg State University, St. Petersburg, Russia
Wiebe Kooistra	Stazione Zoologica Anton Dohrn, Naples, Italy
Barry Leadbeater	University of Birmingham, Birmingham, UK
Julius Lukeš	University of South Bohemia, České Budějovice, Czech Republic
Denis Lynn	University of Guelph, Guelph, Canada
Ramon Massana	Institut de Ciències del Mar, Barcelona, Spain
Edward Mitchell	University of Neuchâtel, Neuchâtel, Switzerland
Jan Pawlowski	University of Geneva, Geneva, Switzerland
Ian Probert	CNRS - Station Biologique de Roscoff, Roscoff, France
Daniel Richter	CNRS - Station Biologique de Roscoff, Roscoff, France
Iñaki Ruiz-Trillo	University of Barcelona, Barcelona, Spain
Gary Saunders	University of New Brunswick, Fredericton, Canada
Raffaele Siano	Ifremer Centre Bretagne, Plouzané, France
Alastair Simpson	Dalhousie University, Halifax, Canada
Pavel Skaloud	Charles University, Prague, Czech Republic
Alexey Smirnov	St. Petersburg State University, St. Petersburg, Russia
Frederick Spiegel	University of Arkansas, Fayetteville AR, USA