EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotic life

Daniel J. Richter[1], Cédric Berney[2,3], Jürgen F. H. Strassert[4], Fabien Burki[5], Colomban de Vargas[2,3]

1. Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta 37-49, 08003 Barcelona, Catalonia, Spain
2. Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR 7144, ECOMAP, 29680 Roscoff, France
3. Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara GOSEE, Paris, France
4. Institute of Biology, Free University of Berlin, Königin-Luise-Straße 1-3, 14195 Berlin, Germany
5. Science for Life Laboratory and Department of Organismal Biology, Uppsala University, Norbyvägen 18D, Uppsala 75236, Sweden

Database Availability

https://doi.org/10.6084/m9.figshare.12417881.v2

Abstract

EukProt is a database of published and publicly available predicted protein sets and unannotated genomes selected to represent eukaryotic diversity, including 742 species from all major supergroups as well as orphan taxa. The goal of the database is to provide a single, convenient resource for studies in phylogenomics, gene family evolution, and other gene-based research across the spectrum of eukaryotic life. Each species is placed within the UniEuk taxonomic framework in order to facilitate downstream analyses, and each data set is associated with a unique, persistent identifier to facilitate comparison and replication among analyses. The database is currently in version 2, and all versions will be permanently stored and made available via FigShare. We invite the community to provide suggestions for new data sets and new annotation features to be included in subsequent versions, with the goal of building a collaborative resource that will promote research to understand eukaryotic diversity and diversification.

Introduction

Over the past 15 years, the discovery of diverse novel microbial eukaryotes, coupled with methods to reconstruct phylogenies based on hundreds of protein-coding genes (known as phylogenomics (Eisen, 2003)) have led to a remarkable reshaping in our understanding of the eukaryotic tree of life, the creation of new supergroups and the placement of enigmatic lineages in known supergroups (Brown et al., 2018; Burki et al., 2012, 2020; Gawryluk et al., 2019; Janouškovec et al., 2017; Kamikawa et al., 2014; Lax et al., 2018; Strassert et al., 2019; Yabuki et al., 2015). The phylogenomic approach has also been used to investigate branching patterns within eukaryotic supergroups, including those with implications for the early evolutionary events in the three major eukaryotic lineages: land plants (Wickett et al., 2014), animals (King & Rokas, 2017) and fungi (Kiss et al., 2019). Furthermore, analyses in diverse eukaryotes have the potential to reveal new genes, pathways and mechanisms of function for biological processes currently characterized only in these three well-studied lineages and their parasites (del Campo et al., 2014; Richter & Levin, 2019). In the ocean alone, planetary-scale metagenomics studies across the full spectrum of life from viruses to animals (Bork et al., 2015; Tara Oceans Coordinators et al., 2020) have already unveiled an extreme diversity

of eukaryotic genes (Carradec et al., 2018) whose phylogenetic origin and ecological function are mostly unknown.

A critical prerequisite to all of these studies is the underlying database of predicted proteins from which orthologs are extracted or other sequence analyses are performed. Because no single website stores the protein predictions for the complete set of eukaryotic taxa that have been sequenced at a genomic scale, each study has needed to assemble their own set, reducing reproducibility among analyses (due to the inclusion of different taxa, or different data sets for the same taxon), and producing a significant barrier to new researchers entering the field. In addition, because the large majority of the protein data sets from diverse eukaryotes are not included in major databases (see Table 1), researchers cannot easily access them via standard tools such as NCBI BLAST (Sayers et al., 2020) in order to find diverse eukaryotic homologs of a protein of interest, nor are valuable annotations such as protein domains (e.g., Pfam (El-Gebali et al., 2019), Interpro (Mitchell et al., 2019)) or gene ontology (The Gene Ontology Consortium, 2019) easily accessible for searching.

To address this gap, we gathered the predicted proteins from a comprehensive set of species representing known eukaryotic diversity. We placed each species within a unified taxonomic framework, UniEuk (Berney et al., 2017), in order to ensure that the evolutionary relationships among data sets are accurately and consistently described. Our database is designed to prioritize ease of use, with unique, persistent identifiers assigned to each data set and a standard system of nomenclature to facilitate repeatability of analyses.

The EukProt Database

The current version of EukProt (version 2) contains 742 eukaryotic species from 5 different types of sequence data (Figure 1): genome (240 species), single-cell genome (25), transcriptome (453), single-cell transcriptome (7) and expressed sequence tag (EST; 17). The data sets were downloaded from 30 different sources (Table 1), with the two principal sources being NCBI (Sayers et al., 2020) and the Moore Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al., 2014). The representation among eukaryotic lineages is highly uneven, which is due to the difficulty of discovering, culturing or sequencing species from many lineages, as well as a bias towards sequencing either macroscopic, multicellular species or unicellular species that are parasites (e.g., Apicomplexa), photosynthetic (such as diatoms, which are part of the Ochrophyte lineage), or are otherwise economically important (del Campo et al., 2014).

The EukProt database is organized around 5 guiding principles:

Breadth of species phylogenetic diversity: the objective of the database is to represent known eukaryotic diversity as fully as possible. In practice, this means that all species with available data sets are included for the majority of lineages. The exceptions are the land plants (Streptophyta), animals (Metazoa) and fungi, and their parasites/pathogens (members of Apicomplexa, Peronosporomycetes, Metakinetoplastina and Fornicata), for which we included a subset of species selected to represent phylogenetic diversity within each group. Within animals, we also emphasized the inclusion of marine planktonic species, as we anticipate these could serve as mapping targets for data from large-scale metagenomic and metatranscriptomic ocean sequencing projects (e.g., *Tara* Oceans (Carradec et al., 2018; Karsenti et al., 2011; Richter et al., 2019)).

Convenience of access: the full data set and its associated metadata can be downloaded from FigShare with a single click. For 501 of the 724 species in the database, publicly available protein sequences were ready to be included directly. For the remaining species, we processed the publicly available data to produce protein sequences for inclusion in the database. The most common actions

were merging independent sets of protein predictions for the same species (105 species), translating mRNA sequences from transcriptomes or EST projects (67), and *de novo* assembling and translating transcriptomes from raw read data (47; we also provide these assemblies as part of the database, as they are not publicly available). A full list of the different types of actions is described in the Methods, and the actions taken for each species are available in the database metadata. We note that 16 species in the database are represented by genomic data lacking accompanying protein annotations; 15 of these are single-cell genomes. We considered annotation of these genomes to be outside the scope of this release, so we include their nucleotide data in a separate file, which can be searched for proteins of interest using translated homology search software (e.g., tblastn (Altschul, 1997) or BLAT (Kent, 2002)).

A unified taxonomic framework: we placed all species into a unified taxonomic framework, UniEuk (Berney et al., 2017), which is based on the most recent consensus eukaryotic taxonomy (Adl et al., 2019). This will maximize interoperability of the database with other resources built upon to the UniEuk taxonomic framework (e.g., EukBank and EukMap (Berney et al., 2017), EcoTaxa (https://ecotaxa.obs-vlfr.fr), and EukRibo (Berney et al., in prep)), and ensure that results arising from the database will be able to use a common set of identifiers to describe analyses at different taxonomic levels. For example, this might include labeling groups in a phylogenetic tree, or summarizing read placement data by taxonomic group. Knowledge of the taxonomy can also facilitate the detection of mis-identified gene sequences within a data set, via comparison of the topology of gene trees to the topology of the taxonomy. In addition to the full taxonomic lineage, we provide for each species the “supergroup”, a very high-level taxonomic grouping, and “taxogroup”, a more fine-grained grouping into evolutionarily or ecologically relevant lineages (see Methods). Finally, for each species we include a list of names it was previously known by, to prevent issues resulting from revisions to species names or lineages, which have occurred frequently with improvements in sequencing and phylogenetic techniques and the discovery of new eukaryotic organisms.

Appropriate references to publications and data sources: to ensure that the researchers who generated and provided each data set receive appropriate credit, we provide the DOI of the publication describing each data set as well as the URL from which it was downloaded. The list of URLs should also allow users of the database to download the original sequences for each data set (although it is not possible to guarantee that the URLs for all data providers will be permanently available).

Reusability, persistence and replicability: the database will be released in successive versions, each of which will be permanently stored and accessible at FigShare. In that way, analyses using the database will need only to specify which version was used, enabling follow-up analyses or replications to begin with the identical database. In addition, each individual data set within the database is assigned a unique, permanent identifier. When a new data set becomes available for a given species, it is assigned a new unique identifier (and the identifier of the data set it replaces, if any, is indicated in the database metadata). These and all other changes between versions of the database will be logged as appropriate.

Growing the EukProt Database with Community Involvement

The core functionality of the database is the distribution of genome-scale protein sequences across the diversity of eukaryotic life and within the UniEuk framework. However, we anticipate that numerous other features might be useful to the community, and we hope to involve the community in suggesting and adding new features in successive versions. These may include information or analyses on full data sets, such as the sequencing technology that was used (e.g., Illumina, PacBio), the 18S ribosomal DNA sequence corresponding to the data set, the completeness of the data set as

estimated by software such as BUSCO (Simão et al., 2015), or the estimation of potential contamination levels with non-target species, as inferred using systematic sequence homology searches. Additionally, we could consider adding features on the level of individual protein sequences, such as protein domains from Pfam (El-Gebali et al., 2019)/Interpro (Mitchell et al., 2019), gene ontology (The Gene Ontology Consortium, 2019)/eggNOG (Huerta-Cepas et al., 2019), or the assignment of unique identifiers to individual protein sequences (currently, FASTA files are distributed as is, without modification to headers/identifiers).

As new genome-scale eukaryotic protein data sets become available, we plan to add them to the database. As yet, we do not have a formal mechanism to accomplish this, and will instead depend on monitoring the literature and assistance from the community. We also plan to add any data sets we may have inadvertently overlooked when building the current version of the database.

In the longer term, we hope the standardization of our database provides a path towards including all data sets in a major sequence repository such as NCBI/EBI/DDBJ, so that they can be more broadly accessible and integrated into the suites of tools available at these repositories.

Methods

Species and strain identity

We determined species and strain identities by reading the publications that described the data sets, consulting the literature for naming revisions, and comparing 18S ribosomal DNA sequences to reference sequence databases. For species that were previously known by other names, we recorded them in the metadata for the data set, except in cases where a species was originally assigned to a genus but not identified to the species level (e.g., *Goniomonas* sp., now identified as *Goniomonas avonlea*, is not listed as a previous name).

Supergroups and taxogroups from UniEuk

Unlike the taxonomic system used in some other resources, the full taxonomic lineages we provide for all species (which follow the framework developed in the UniEuk project) are not based on a fixed number of ranks, but on a free, unlimited number of taxonomic levels, in order to match phylogenetic evidence as closely as possible. This provides end-users more information and flexibility, but could also make it more difficult to summarize results of downstream analyses. Therefore, we provide two additional fields (“supergroup” and “taxogroup”) to help end-users whenever it is useful to distribute eukaryotic diversity into a fixed number of taxonomic categories of equivalent phylogenetic depth or ecological relevance. The 42 “supergroups” (of which 38 are included in EukProt) consist of strictly monophyletic, deep-branching eukaryotic lineages of a phylogenetic depth equivalent to the “classic” Alveolata, Rhizaria, and Stramenopiles. They are therefore highly variable in relative diversity, ranging from clades consisting of a single, orphan genus (e.g. *Ancoracysta*, *Mantamonas*, *Palpitomonas*), to Metazoa as a whole. The “taxogroup” level allows further subdivision of large supergroups into lineages of relatively equivalent evolutionary or ecological relevance, based on current knowledge. This level is more arbitrarily defined; it can include paraphyletic groupings in highly diversified supergroups to accommodate minor lineages of similar ecology or phylogenetic depth. As illustrative examples, diatoms are in the Diatomeae taxogroup within the Stramenopiles supergroup, and coccolithophores are in the Prymnesiophyceae taxogroup within the Haptophyta supergroup. Small, ecologically and morphologically homogeneous supergroups are not subdivided further; in such cases the “taxogroup” level is the same as the “supergroup” level. The same approach will be used in

EukRibo, a database of reference ribosomal RNA gene sequences developed in parallel (Berney et al., in prep) to help users link analyses of different types of genetic data.

Merging strains from the same species

In general, we only included data from a single strain/isolate per species. However, when only a single transcriptome data set was available for a given strain of a species, and there were additional published transcriptome data sets for other strains of the same species, we combined them using CD-HIT (Li & Godzik, 2006) run with default parameter values, in order to guard against the possibility that a single transcriptome might lack genes expressed only in one condition or experiment. When multiple strains were merged to produce a species' data set (there were 25 such cases), this information is indicated in the metadata for the data set.

Processing steps applied to publicly available data

EukProt metadata indicate the additional steps applied to each data set after downloading from the data source (if any). All software parameter values were default, unless otherwise specified below or in the metadata record for a given data set.

'assemble mRNA': *de novo* transcriptome assembly using Trinity v. 2.8.4, http://trinityrnaseq.github.io/ (Haas et al., 2013). We trimmed Illumina input reads for adapters and sequence quality using the built-in '--trimmomatic' option. We trimmed 454 input reads prior to running Trinity with Trimmomatic v. 0.3.9, http://www.usadellab.org/cms/?page=trimmomatic (Bolger et al., 2014) with the directives 'ILLUMINACLIP:[454 adapters FASTA file]:2:30:10 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25'.

'translate mRNA': *de novo* translation of mRNA sequences with Transdecoder v. 5.3.0, http://transdecoder.github.io/. When the number of predicted protein sequences for a given species was less than half of the input mRNA sequences, we reduced the minimum predicted protein length to 50 (from the default of 100).

'CD-HIT': clustering of protein sequences to produce a non-redundant data set using CD-HIT v. 4.6, http://weizhongli-lab.org/cd-hit/ (Li & Godzik, 2006). We used this tool principally to combine protein predictions for different strains of the same species, but also to reduce the size of very large predicted protein sets (> 50,000 proteins) that showed evidence of redundancy.

'extractfeat', 'transeq', and 'trimseq': from the EMBOSS package v. 6.6.0.0, http://emboss.sourceforge.net/ (Rice et al., 2000). We used extractfeat to produce coding sequences (CDS) from genomes with gene annotations but without publicly available protein sequences. We used transeq to translate CDS directly into proteins. We used trimseq to trim EST sequences before translation with Transdecoder.

The EukProt database distribution on FigShare

The database is distributed in a single archive containing five files. One file contains 726 protein data sets, for species with either a genome (239) or single-cell genome (10) with predicted proteins, a transcriptome (453), a single-cell transcriptome (7), or an EST assembly (17). A second file contains 16 genomes lacking predicted protein annotations, for 15 species with single-cell genomes, and 1 species with a genome sequence. These can be queried for proteins of interest with translated sequence homology search software. A third file contains assembled transcriptome contigs, for 53 species with publicly available mRNA sequence reads but no publicly available assembly. The proteins predicted from these assemblies are included in the proteins file. Finally, the database

metadata are distributed as two files: one file for the data sets included in the current version of the database (742), and a second file for data sets not included (50), accompanied by the reason they were not included (for example, if the sequences are published but not publicly available or if they were replaced by a higher-quality data set for the same species).

Acknowledgements

We thank Michelle Leger for helpful advice, Núria Ros i Rocher for suggestions on the figure, and members of the Multicellgenome Lab for discussions. This work was supported by the French Government 'Investissement d'Avenir' program OCEANOMICS (ANR-11-BTBR-0008) and the ABiMS computing cluster at the Station Biologique de Roscoff, France. DJR was supported by postdoctoral fellowships from the Beatriu de Pinós programme of the Government of Catalonia's Secretariat for Universities and Research of the Ministry of Economy and Knowledge, and from "la Caixa" Foundation (ID 100010434), with the fellowship code LCF/BQ/PI19/11690008. CB was supported by a grant from the Gordon and Betty Moore Foundation (GBMF5257 / UniEuk) and is grateful to the International Society of Protistologists for additional support.

References

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., del Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., … Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4–119. https://doi.org/10.1111/jeu.12691

Altschul, S. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Berney, C., Ciuprina, A., Bender, S., Brodie, J., Edgcomb, V., Kim, E., Rajan, J., Parfrey, L. W., Adl, S., Audic, S., Bass, D., Caron, D. A., Cochrane, G., Czech, L., Dunthorn, M., Geisen, S., Glöckner, F. O., Mahé, F., Quast, C., … de Vargas, C. (2017). *UniEuk*: Time to Speak a Common Language in Protistology! *Journal of Eukaryotic Microbiology*, *64*(3), 407–411. https://doi.org/10.1111/jeu.12414

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., & Wincker, P. (2015). Tara Oceans studies plankton at planetary scale. *Science*, *348*(6237), 873–873. https://doi.org/10.1126/science.aac5605

Brown, M. W., Heiss, A. A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., Shiratori, T., Ishida, K.-I., Hashimoto, T., Simpson, A. G. B., & Roger, A. J. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biology and Evolution*, *10*(2), 427–433. https://doi.org/10.1093/gbe/evy014

Burki, F., Okamoto, N., Pombert, J.-F., & Keeling, P. J. (2012). The evolutionary history of haptophytes and cryptophytes: Phylogenomic evidence for separate origins. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1736), 2246–2254. https://doi.org/10.1098/rspb.2011.2301

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology & Evolution*, *35*(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez,

G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1), 373. https://doi.org/10.1038/s41467-017-02342-1

del Campo, J., Sieracki, M. E., Molestina, R., Keeling, P., Massana, R., & Ruiz-Trillo, I. (2014). The others: Our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, *29*(5), 252–259. https://doi.org/10.1016/j.tree.2014.03.006

Eisen, J. A. (2003). Phylogenomics: Intersection of Evolution and Genomics. *Science*, *300*(5626), 1706–1707. https://doi.org/10.1126/science.1086292

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Gawryluk, R. M. R., Tikhonenkov, D. V., Hehenberger, E., Husnik, F., Mylnikov, A. P., & Keeling, P. J. (2019). Non-photosynthetic predators are sister to red algae. *Nature*, *572*(7768), 240–243. https://doi.org/10.1038/s41586-019-1398-6

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Janouškovec, J., Tikhonenkov, D. V., Burki, F., Howe, A. T., Rohwer, F. L., Mylnikov, A. P., & Keeling, P. J. (2017). A New Lineage of Eukaryotes Illuminates Early Mitochondrial Genome Reduction. *Current Biology*, *27*(23), 3717-3724.e5. https://doi.org/10.1016/j.cub.2017.10.051

Kamikawa, R., Kolisko, M., Nishimura, Y., Yabuki, A., Brown, M. W., Ishikawa, S. A., Ishida, K., Roger, A. J., Hashimoto, T., & Inagaki, Y. (2014). Gene Content Evolution in Discobid Mitochondria Deduced from the Phylogenetic Position and Complete Mitochondrial Genome of Tsukubamonas globosa. *Genome Biology and Evolution*, *6*(2), 306–315. https://doi.org/10.1093/gbe/evu015

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., … Wincker, P. (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biology*, *9*(10), e1001177. https://doi.org/10.1371/journal.pbio.1001177

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology*, *12*(6), e1001889. https://doi.org/10.1371/journal.pbio.1001889

Kent, W. J. (2002). BLAT—The BLAST-Like Alignment Tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

King, N., & Rokas, A. (2017). Embracing Uncertainty in Reconstructing Early Animal Evolution. *Current Biology*, *27*(19), R1081–R1088. https://doi.org/10.1016/j.cub.2017.08.054

Kiss, E., Hegedüs, B., Virágh, M., Varga, T., Merényi, Z., Kószó, T., Bálint, B., Prasanna, A. N., Krizsán, K., Kocsubé, S., Riquelme, M., Takeshita, N., & Nagy, L. G. (2019). Comparative

genomics reveals the origin of fungal hyphae and multicellularity. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-12085-w

Lax, G., Eglit, Y., Eme, L., Bertrand, E. M., Roger, A. J., & Simpson, A. G. B. (2018). Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature*, *564*(7736), 410–414. https://doi.org/10.1038/s41586-018-0708-8

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research*, *47*(W1), W256–W259. https://doi.org/10.1093/nar/gkz239

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Marron, A. O., Ratcliffe, S., Wheeler, G. L., Goldstein, R. E., King, N., Not, F., de Vargas, C., & Richter, D. J. (2016). The Evolution of Silicon Transport in Eukaryotes. *Molecular Biology and Evolution*, *33*(12), 3226–3248. https://doi.org/10.1093/molbev/msw209

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., … Finn, R. D. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, *47*(D1), D351–D360. https://doi.org/10.1093/nar/gky1100

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Richter, D. J., & Levin, T. C. (2019). The origin and evolution of cell-intrinsic antibacterial defenses in eukaryotes. *Current Opinion in Genetics & Development*, *58–59*, 111–122. https://doi.org/10.1016/j.gde.2019.09.002

Richter, D. J., Watteaux, R., Vannier, T., Leconte, J., Frémont, P., Reygondeau, G., Maillet, N., Henry, N., Benoit, G., Fernàndez-Guerra, A., Suweis, S., Narci, R., Berney, C., Eveillard, D., Gavory, F., Guidi, L., Labadie, K., Mahieu, E., Poulain, J., … Tara Oceans Coordinators. (2019). *Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems* [Preprint]. Ecology. https://doi.org/10.1101/867739

Sayers, E. W., Beck, J., Brister, J. R., Bolton, E. E., Canese, K., Comeau, D. C., Funk, K., Ketter, A., Kim, S., Kimchi, A., Kitts, P. A., Kuznetsov, A., Lathrop, S., Lu, Z., McGarvey, K., Madden, T. L., Murphy, T. D., O'Leary, N., Phan, L., … Ostell, J. (2020). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *48*(D1), D9–D16. https://doi.org/10.1093/nar/gkz899

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V., & Burki, F. (2019). New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. *Molecular Biology and Evolution*, *36*(4), 757–765. https://doi.org/10.1093/molbev/msz012

Tara Oceans Coordinators, Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M. B., Wincker, P., & de Vargas, C. (2020). Tara Oceans: Towards global ocean ecosystems biology. *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-020-0364-5

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., … Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification

of land plants. *Proceedings of the National Academy of Sciences*, *111*(45), E4859–E4868. https://doi.org/10.1073/pnas.1323926111

Wideman, J. G., Lax, G., Leonard, G., Milner, D. S., Rodríguez-Martínez, R., Simpson, A. G. B., & Richards, T. A. (2019). A single-cell genome reveals diplonemid-like ancestry of kinetoplastid mitochondrial gene structure. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1786), 20190100. https://doi.org/10.1098/rstb.2019.0100

Yabuki, A., Kamikawa, R., Ishikawa, S. A., Kolisko, M., Kim, E., Tanabe, A. S., Kume, K., Ishida, K., & Inagki, Y. (2015). Palpitomonas bilix represents a basal cryptist lineage: Insight into the character evolution in Cryptista. *Scientific Reports*, *4*(1). https://doi.org/10.1038/srep04641

426 **Table 1.** Web hosts from which data sets were downloaded. The count for figshare.com includes 314
427 species from the MMETSP (Keeling et al., 2014) for which a procedure to remove cross-
428 contamination was applied (Marron et al., 2016).
429

Hostname	Number of Data Sets
figshare.com	351
ncbi.nlm.nih.gov	281
datadryad.org	32
ebi.ac.uk	16
rutgers.edu	9
genoscope.cns.fr	8
yadi.sk	6
oist.jp	5
jgi.doe.gov	3
genomics.cn	3
compagen.org	2
nrifs.fra.affrc.go.jp	2
obs-vlfr.fr	1
zenodo.org	1
licebase.org	1
ufl.edu	1
drive.google.com	1
liv.ac.uk	1
bioenergychina.org	1
algaegenome.org	1
giardiadb.org	1
nal.usda.gov	1
ovgu.de	1
umontreal.ca	1
ugent.be	1
bitbucket.org	1
davidadlergold.com	1
sciencemag.org	1
malab.cn	1
treegenesdb.org	1

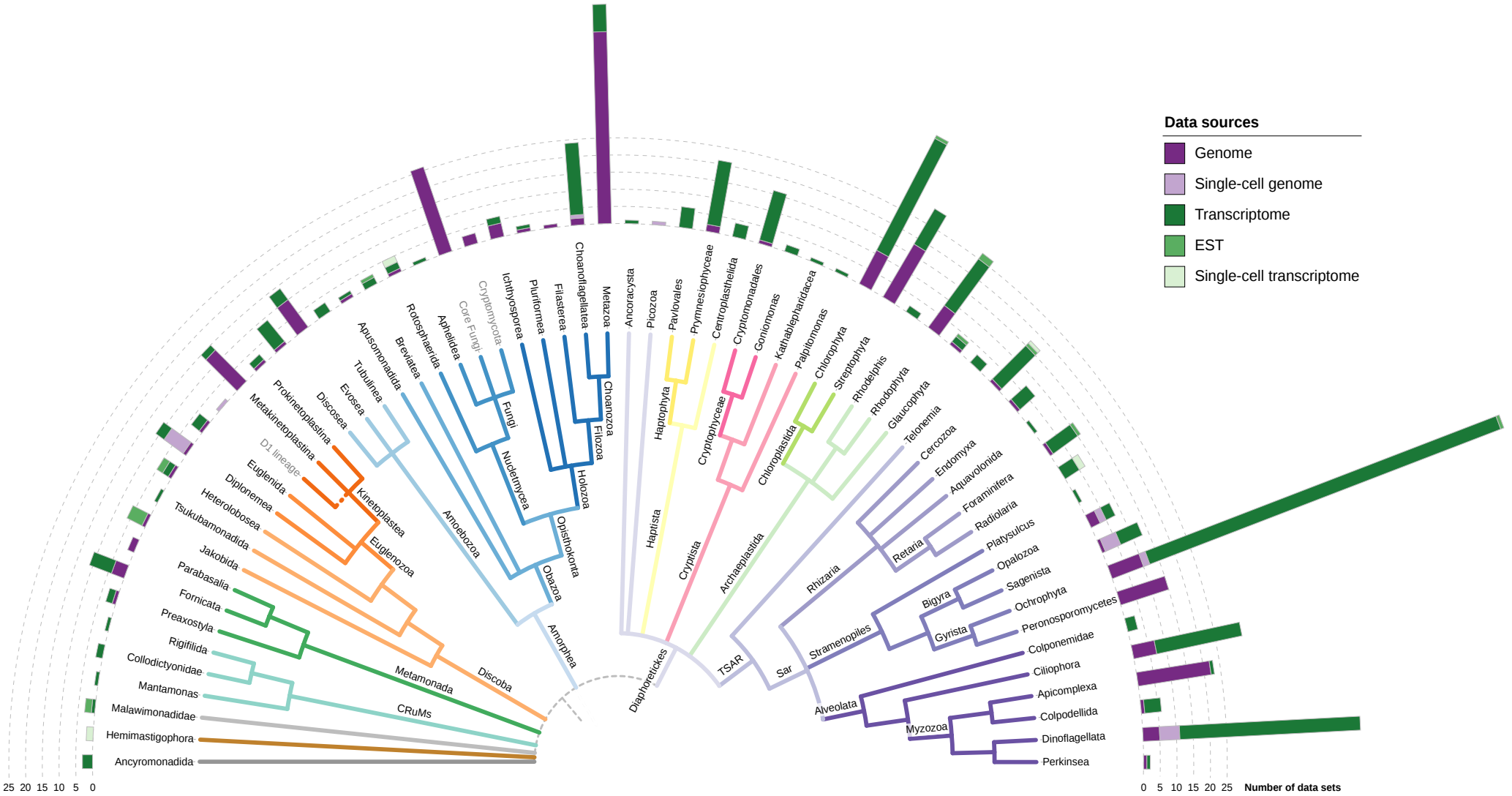

Figure 1. Distribution of 742 source data sets on the eukaryotic tree of life, separated by data set type, with taxonomy based on UniEuk (Adl et al., 2019; Berney et al., 2017). The position of the eukaryotic root, indicated with a dashed line, is currently unresolved. Group names shown in grey are not officially recognized: “D1 lineage” refers to a single-cell genome most closely related to Kinetoplastea (Wideman et al., 2019), whose affiliation with the group is indicated with a dashed line; “Cryptomycota” represents rozellids and microsporidia; “Core Fungi” encompasses all other fungal lineages from chytrids to Dikarya. The figure was created with iTOL (Letunic & Bork, 2019).