CellPress
OPEN ACCESS

Trends in
Genetics

Review

Protist genomics: key to understanding eukaryotic evolution

Alexandra Schoenle[1,2,*], Ore Francis[3], John M. Archibald[4,5], Fabien Burki[6], Jan de Vries[7], Kenneth Dumack[8], Laura Eme[9], Isabelle Florent[10], Elisabeth Hehenberger[11], Tarja T. Hoffmeyer[12], Iker Irisarri[13], Enrique Lara[14], Michelle M. Leger[15], Julius Lukeš[11,16], Ramon Massana[17], Varsha Mathur[18], Frank Nitsche[19], Jürgen F.H. Strassert[20,24], Alexandra Z. Worden[21], Vyacheslav Yurchenko[22], Javier del Campo[23,25], and Ann-Marie Waldvogel[1,2,25]

All eukaryotes other than animals, plants, and fungi are protists. Protists are highly diverse and found in nearly all environments, with key roles in planetary health and biogeochemical cycles. They represent the majority of eukaryotic diversity, making them essential for understanding eukaryotic evolution. However, these mainly unicellular, microscopic organisms are understudied and the generation of protist genomes lags far behind most multicellular lineages. Current genomic methods, which are primarily designed for animals and plants, are poorly suited for protists. Advancing protist genome research requires reevaluating plant- and animal-centric genomic standards. Future efforts must leverage emerging technologies and bioinformatics tools, ultimately enhancing our understanding of eukaryotic molecular and cell biology, ecology, and evolution.

Highlights

Protists comprise most of eukaryotic genomic and cellular diversity but are the least studied eukaryotes. Genome sequencing from diverse protist taxa is crucial for understanding eukaryote evolution.

Large genome size variation and complex gene networks in protists pose challenges and opportunities for genomics and bioinformatics.

Current difficulties with culturing and sequencing heterotrophic and symbiotic species highlight the need for targeted technological advancements.

Developments in single-cell genomics, metagenomics, and long-read sequencing make it possible to study rare and uncultured protists.

Advances in protist genomics will depend on creating group-specific methodologies considering their complexity across levels of diversity, with the long-term goal of generating high-quality genomes of unicellular eukaryotes.

Protists in the genomic era

Comprehensive genomic resources are crucial for understanding the biological diversity and molecular function of organisms. The spin-offs from this knowledge have implications for many fields of study, from biogeochemistry to medicine and conservation biology, to name but a few [1,2]. Hence, there has been a marked increase in the availability of genomic data across various taxonomic groups, in part driven by several international genome initiatives [e.g., **Earth BioGenome Project (EBP)**, **Darwin Tree of Life (DToL)**, **European Reference Genome Atlas (ERGA)**; see Glossary], which aim to catalog and characterize the genomes of Earth's eukaryotic biodiversity [3,4]. Recent technological advances provide strategies for generating chromosome-scale reference genomes for many disparate organisms across the Tree of Life [5,6]. Despite this progress, **protists** (Box 1) remain under-represented in genomic research. Once viewed as 'simple organisms', advances in microscopy, transcriptomics, and comparative genomics have challenged this view, revealing their genetic diversity and complex evolution within the Tree of Life [7,8] (Box 1). Prioritizing protist genomics within genome initiatives, such as the Aquatic Symbiosis Genomics Project [9], is essential for enhancing our understanding of their diversity, biology, and ecology, and for contextualizing and advancing knowledge of fungi, plants, and animals [10–12].

The role of protist genomes in decoding eukaryotic evolution

Reconstructing the eukaryotic tree of life

Protists represent the entirety of phylogenetic diversity of eukaryotes [13], making our understanding of eukaryotic evolution heavily dependent on how these protist lineages are related to each other. At first, phylogenetic trees were based on comparisons of morphology and metabolism, but beginning in the 1970s, the field was revolutionized by the use of molecular sequence

[1]Ecological Genomics, Department of Biology, Institute of Zoology, Biocenter Cologne, University of Cologne, Cologne, Germany
[2]Global Change Limnology, School of Life Sciences, Technical University of Munich, Munich, Germany
[3]Wellcome Sanger Institute, Cambridge, UK
[4]Department of Biochemistry & Molecular Biology, Dalhousie University, Halifax, Canada
[5]Institute for Comparative Genomics, Dalhousie University, Halifax, Canada

 https://doi.org/10.1016/j.tig.2025.05.004

data to infer evolutionary relationships. However, single-gene markers (e.g., rRNA genes) often fail to recover deep eukaryotic relationships (e.g., because of too variable/conserved regions) [14,15]. Phylogenomic approaches, using hundreds to thousands of loci and protein-coding genes, are less prone to these biases. These have led to major insights on topics such as the acquisition/reduction of plastids and mitochondria, the origin of animals, and even the root of eukaryotes [16–19]. Incorporating more genomic and transcriptomic data from diverse eukaryotes – especially protists – is required to improve our understanding of eukaryotic evolution and reveal previously hidden evolutionary trajectories.

Understanding transitions in eukaryotic evolution

Symbiosis is one of the most influential aspects of the rise of eukaryotic life and evolution. Protists exhibit a remarkable range of symbiotic interactions with other life forms, such as bacteria, archaea, and animals. Among these diverse symbiotic interactions, parasitism (a symbiosis in which one of the partners negatively impacts the fitness of the other) represents one specialized form of symbiosis, and most of the available protist genomes correspond to human or livestock parasites [12], often associated with extreme genomic adaptations and host dependency. Examining shifts of symbiotic partnerships can help understand key evolutionary transitions, such as the transformation from mutualistic relationships to parasitic ones, offering insights into the principles and dynamics of coevolution. The genome analysis of the free-living counterparts of human pathogenic kinetoplastids, apicomplexans, and holomycotans revealed that each group independently evolved intracellular parasitism by acquiring different adaptations. However, certain common themes are shared among the unrelated groups. These themes include, for example, flagellar loss, expansion of transportation gene families, and genome rearrangements facilitating diversification and accelerated proliferation [20,21].

Endosymbiosis has significantly impacted eukaryotic diversification. In addition to countless modern endosymbioses between protists and other protists, animals, and prokaryotes, the merging of one cell into another gave rise to defining **organelles** of the eukaryotic cell – mitochondria and plastids [22,23]. Gene transfer (from the organelles) to the host nucleus has contributed to drastic genome size reduction in plastids and mitochondria, rendering the function of these compartments dependent on the import of nucleus-encoded proteins. Our understanding of the origin and evolution of these organelles has advanced significantly since the hypothesis of their endosymbiotic origin was first proposed – including increasingly blurred boundaries between what we initially defined as endosymbiont and what as organelle – much of which has been informed by genomic data [24]. More organellar and nuclear genomes are needed to further refine models of organelle endosymbiosis at the most fundamental levels – such as identifying the interacting partners and understanding the sequence of steps required to transform an endosymbiont into a true organelle.

Eukaryogensis and the evolution of the eukaryotic cell structures are also a key evolutionary event that can be better understood thanks to the comparative analysis of protist genomes and their prokaryotic relatives – archea and bacteria [25,26].

Eukaryotic multicellularity evolved multiple times independently in plants, animals, fungi, and algae, alongside simpler forms of clonal multicellularity [27,28]. Genomic data from protists has provided at least two key insights into multicellularity. First, phylogenomic analyses revealed evolutionary patterns, including simple clonal and/or aggregative stages in protists related to complex multicellular organisms [29–31], and back-and-forth transitions between unicellular and multicellular states in streptophyte algae [32,33] and fungi [34]. Second, comparative genomics showed that while the molecular machinery underpinning cell differentiation, adhesion, and

[6]Department of Organismal Biology, Uppsala University, Uppsala, Sweden
[7]University of Göttingen, Institute of Microbiology and Genetics, Göttingen Center for Molecular Biosciences (GZMB), Campus Institute Data Science (CIDAS), Dept. of Applied Bioinformatics, 37077 Göttingen, Germany
[8]Terrestrial Ecology, Department of Biology, Institute of Zoology, Biocenter Cologne, University of Cologne, Cologne, Germany
[9]Unité d'Ecologie Systématique et Evolution, CNRS, AgroParisTech, Université Paris-Saclay, Gif Sur Yvette, France
[10]UMR7245, National Museum of Natural History and CNRS, Paris, France
[11]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czechia
[12]Department of Biology, Institute of Zoology, Biocenter Cologne, University of Cologne, Cologne, Germany
[13]Department of Biodiversity and Evolutionary Biology, Museo Nacional de Ciencias Naturales (MNCN-CSIC), c/ José Gutiérrez Abascal 2, 28006 Madrid, Spain
[14]Department of Mycology, Royal Botanical Garden-CSIC, Madrid, Spain
[15]Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan
[16]Faculty of Science, University of South Bohemia, České Budějovice, Czechia
[17]Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM-CSIC), Barcelona, Catalonia, Spain
[18]Faculty of Life Sciences, University of Vienna, Vienna, Austria
[19]General Ecology, Department of Biology, Institute of Zoology, Biocenter Cologne, University of Cologne, Cologne, Germany
[20]Department of Evolutionary and Integrative Ecology, Leibniz Institute of Freshwater Ecology and Inland Fisheries, Berlin, Germany
[21]Marine Biological Laboratory, Woods Hole, MA, USA
[22]Life Science Research Centre, University of Ostrava, Ostrava, Czechia
[23]Biodiversity Program, Institut de Biologia Evolutiva (CSIC - Universitat Pompeu Fabra), Barcelona, Catalonia, Spain
[24]Department of Environmental Microbiomics, Institute of Environmental Technology, Technical University of Berlin, Berlin, Germany
[25]These authors share senior authorship

*Correspondence:
a.schoenle@tum.de (A. Schoenle).

communication differs across multicellular lineages [17,35,36], gene family expansion, lateral gene transfer, and domain shuffling were important evolutionary drivers [36–41]. Many genes initially considered to be specific to multicellular organisms have since been found in unicellular relatives, suggesting they predate multicellularity [42–45]. These functionally important genes often show patchy phylogenetic distribution [29,46], highlighting the importance of considering protist diversity to reconstruct ancestral multicellular stages.

Terrestrialization has occurred independently in diverse eukaryotic lineages, including algae and plants, fungi, **ciliates**, rhizarians, or stramenopiles, among others [47]. These transitions involved challenges such as exposure to unbuffered stressors (e.g., drastic temperature changes, higher light, and UV intensities) and water scarcity (osmotic stress). Comparative genomic studies are just beginning to uncover the genomic basis for the terrestrialization of life, foremost in plants by comparing them with their closest (streptophyte) algal ancestors [43,44]. Further studies across the broader diversity of eukaryotic groups will improve our understanding of common (and lineage-specific) adaptations during terrestrialization, for example, in fungi [48].

Revealing the diversity of molecular features and cellular components

Ciliates are well-known for their deviations from the standard genetic code, yet the list of other protist lineages with **noncanonical genetic codes** is expanding fast [49], with advances in the development of genetic tools in diverse organisms [50]. Protists have already served on several occasions as 'eye-openers', with discoveries such as RNA editing, *trans*-splicing, polycistronic transcription, and variant surface proteins, all of which were later found also in various multicellular organisms [51,52]. Unicellular eukaryotes are likely to reveal more novel molecular features, even previously unrecognized organelles, as exemplified by the discovery of '**nitroplasts**' [53]. Sequencing more protist genomes will undoubtedly continue to reveal major departures from what are currently considered typical (and to varying extents, immutable) molecular and biochemical features of the eukaryotic cell.

Decoding ecosystem dynamics

In most natural habitats, protist assemblages are formed by many species belonging to distant taxonomic groups. Here, **metagenomics** and **metatranscriptomics** allow the discovery of their community composition, genomic potential, and gene expression [47,54]. At present, such studies are limited by the current scarcity or discontinuity and incompleteness of reference genomes, hindering the identification of the key players responsible for ecosystem function and health. Integrating high-quality genomes with extensive 'omics' data will make it possible to address key aspects that advance our understanding of protist roles in biogeochemical cycles [55] and community dynamics by ecological interaction networks [56], such as microbial food webs. This will enable tracing allele frequencies and community dynamics, shedding light on how global change affects ecosystem structure and stability.

Particularities of genome research with protists

Although approximately 3000 protist species were culturable as of 2014 [9], representing only 0.03% of the predicted protist diversity (Box 1), many species remain uncultured and belong to previously unrecognized lineages. Thus, different laboratory approaches for generating protist genomes from cultured and uncultured protists are currently used (Box 2). Two emerging culturing-independent approaches are starting to provide protist genomic resources: single-cell genomics [57] and metagenomics [55] (Box 2). Both approaches are in their infancy with regard to their application to protists, but their development is benefiting from associated disciplines such as clinical research. In addition, the species-level diversity and complexity of protists involve

Glossary

Ciliates: unicellular organisms that are characterized by the presence of numerous hair-like structures called cilia, which are used for movement and feeding.
Darwin Tree of Life (DToL): collaborative initiative that aims to sequence the genomes of 70 000 eukaryotic species in Britain and Ireland with the goal of advancing biology, conservation, and biotechnology. Affiliated with EBP. https://www.darwintreeoflife.org/
Earth BioGenome Project (EBP): large multinational, multi-consortium initiative aiming to sequence and catalog the genomes of all currently described eukaryotic species on Earth within a 10-year timeframe. EBP includes initiatives focused on specific taxonomic groups or parts of the world such as Africa, Europe, Taiwan, Chile (start: 2020). https://www.earthbiogenome.org/
European Reference Genome Atlas (ERGA): collaborative initiative across Europe aimed at addressing biodiversity loss by generating high-quality, complete reference genomes for European species. Pan-European partner of EBP. https://www.erga-biodiversity.eu/
Fluorescence-activated cell sorting (FACS): application of flow cytometry to sort a heterogeneous mixture of cells into individual wells or tubes, one cell at a time, based on the specific fluorescence characteristics and light-scattering properties of the target cells. This can be done utilizing various types of fluorescence, including autofluorescence from natural pigmentation or fluorescent labeling.
Internally eliminated sequences (IESs): noncoding DNA regions that are removed from the genome during a process of somatic genome rearrangement, typically after fertilization.
Metagenome-assembled genome (MAG): genome assembled from metagenomic data.
Metagenomics: method used to sequence genomic DNA from multispecies samples.
Metatranscriptomics: method used to sequence RNA from multispecies samples.
Multiple displacement amplification (MDA): method used to amplify genomic DNA through isothermal replication, utilizing a high-fidelity DNA polymerase to generate large amounts

distinct life strategies (e.g., different feeding modes), which must be considered in the lab, requiring specific methodological treatments and adaptations.

Approaches aligned to the life strategies of protists

Many protists are symbionts of animals (e.g., apicomplexans) or plants (e.g., oomycetes), or are themselves hosts to bacterial or other protist endosymbionts (e.g., parabasalids). This complicates their isolation and study as single biological entities. *In vitro* culture methods are largely lacking for such protists and, instead, they rely on field collections of infected hosts, leading to isolates being frequently contaminated with host cells and/or environmental microorganisms. Two strategies can be used to overcome these limitations. The first involves *in vivo* isolation or enrichment of symbiont biomass. This can be achieved by manual isolation using micro-pipetting [58,59], performing differential filtration to remove host tissue [60], or physical separation using density gradient centrifugation [61]. Recently, high-throughput single-cell technologies have begun to complement or replace some of these approaches. These include using **fluorescence-activated cell sorting (FACS)** to isolate cells, followed by **multiple displacement amplification (MDA)** for whole-genome sequencing [62], and microfluidics-based approaches such as the 10x Genomics Chromium system [63]. The second strategy focuses on the *in silico* isolation of symbiont and host genome sequence reads employing k-mer-based approaches [64], GC content, trinucleotide composition, or genome mapping against the host genome when available.

Heterotrophic protists require food, often in the form of bacteria or other protists. This makes the cultivation of such taxa more challenging compared with algae or fungi, resulting in an under-representation of genomic data from heterotrophic protist species. Not all protists can be co-cultured with a single defined food organism; and even when the protist is a bacterivore, not all prey species are equivalent. Moreover, bacteria release secondary metabolites as a defense against predators [65], which selects for protist species with specific resistances. As a result, protist food requirements can differ even between closely-related species [66]. Culturing procedures for radiolarians, which make up to 5% of the total planktonic biomass in surface oceans, are rudimentary at best and aimed simply at maintaining the organisms long enough to be observed; life cycles have never been closed to our knowledge [67]. The situation is similar in foraminiferans, another clade of mostly marine protists, albeit with some recent successes [68]. Many of these large protists (>100 μm) host several symbionts, whose presence is obligatory for protists like dinoflagellates and haptophytes [69].

of DNA from a small or limited sample, often used in applications like single-cell genomics and forensic analysis.
Nitroplast: cyanobacterium-derived nitrogen-fixing organelle that evolved independent of canonical plastids.
Noncanonical genetic code: variations of the genetic code, where the correspondence between codons and amino acids differs from the canonical code. They occur most frequently in particular lineages of protists, mitochondria, and plastids.
Organelle: structure within a eukaryotic cell that has a specific function, analogous to an organ in multicellular organisms (here, used specifically for membrane-bound organelles, e.g., mitochondria, chloroplasts).
Primary assembly: best resolved haplotype genome assembly for a taxon. Alleles that cannot be incorporated into the primary assembly are assembled into an alternate assembly.
Protists: all other eukaryotes that are not fungi, animals, or plants. As a polyphyletic group, they are morphologically and ecologically diverse, mostly microbial, and collectively constitute the majority of eukaryotic life.
Symbiosis: a long-term interaction between two organisms of different species, encompassing a spectrum of relationships from mutualism to commensalism and parasitism.

Box 1. Beginner's guide to protists

Due to their small size, protists were not considered in early theories of eukaryotic relationships.

It was not until 1818 that Goldfuß introduced the taxon 'Protozoa', classifying them as a class of 'simplest' animals (together with 'lower' multicellular forms such as bryozoans and rotifers) [88,89]. In 1866, when photoautotrophic and heterotrophic microbial eukaryotes were unified, Haeckel described the 'Protista' as a sister clade to animals and plants [90]. While further studies using constantly improving microscopy techniques challenged this view as well as the view that protists are proto-types of multicellular organisms [91], and allowed the identification of numerous eukaryotic lineages, which are valid up to the present day; the polyphyletic nature of protists remained enigmatic until the late 20th century.

Protists are highly diverse; they make up the entirety of most eukaryotic phylogenetic diversity within the eukaryotic Tree of Life (Figure I) [92,93], at least two groups of which have been proposed within the last decade alone, while the deeper phylogeny of several groups still remains unresolved [93]. Most protists are unicellular and require microscopy tools and expertise for basic observation. Only a small fraction of protists have thus been scientifically described (76 904 protist species in 2007 [94]) relative to their predicted richness of up to 10.5 million species, with over 1 million species of apicomplexans alone [95–97]. Mass sequencing of molecular markers (e.g., 18S rDNA amplicons) has led to an explosion in the field of phylogenetics, revealing a unique diversity of previously unknown protistan lineages in soil and marine habitats [98,99], which would have been missed with cultivation-based methods. Protists, with cell sizes ranging from <1 μm up to tens of centimeters, are ubiquitously distributed in terrestrial, marine, and freshwater environments [95,100–102], including extreme environments such as the deep sea [103] and the Atacama Desert [104]. They have evolved different behavioral strategies and multiple trophic modes, including autotrophy, heterotrophy, mixotrophy, saprotrophy, parasitism, and a dazzling variety of symbioses [7]. An integrated approach is needed to combine protist culturing with 'omics' approaches, imaging, and high-throughput single-cell manipulation strategies [105] to resolve ecological, evolutionary, and phylogenetic questions arising.

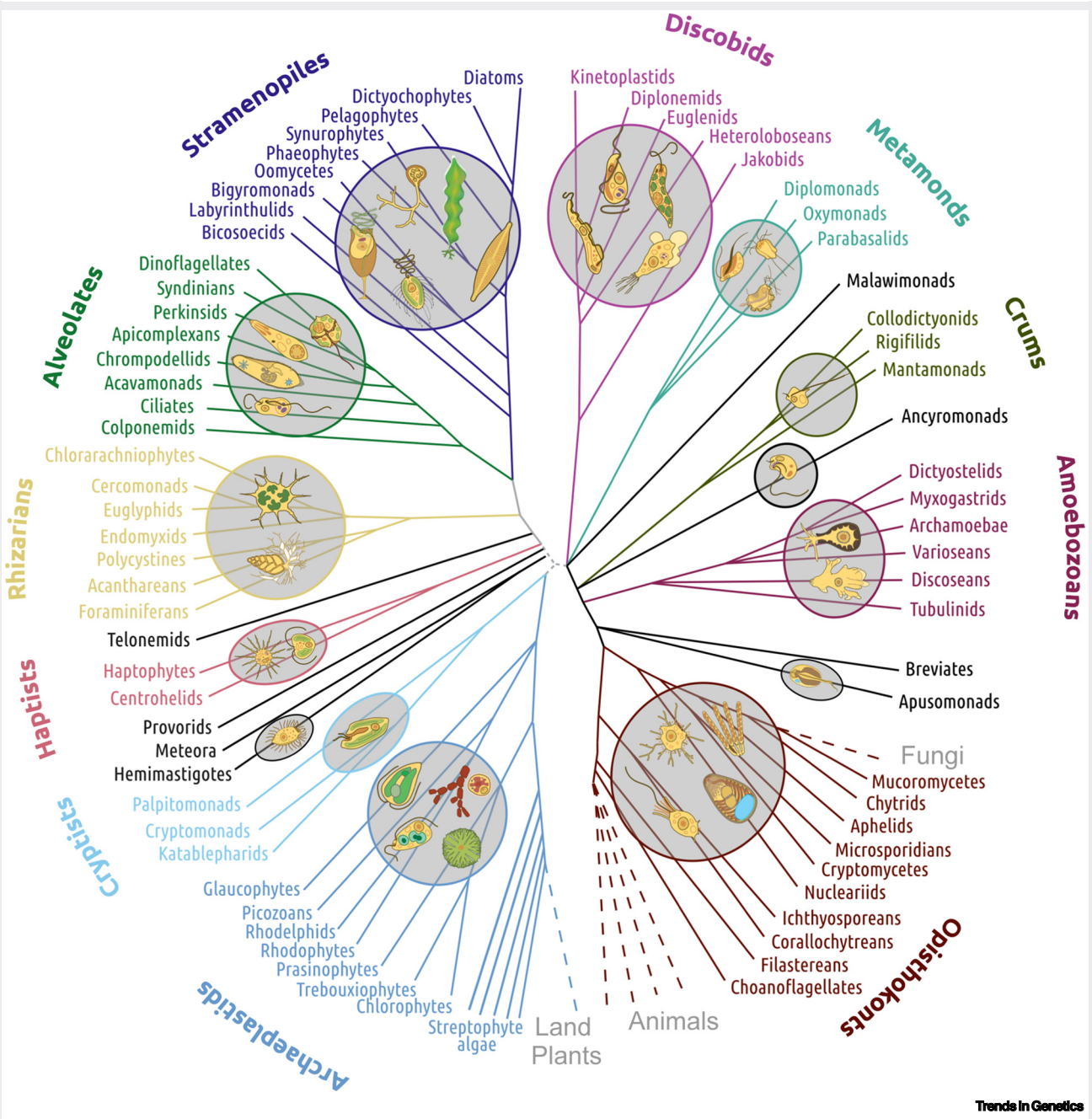


Figure I. The diversity of protists within the eukaryotic tree of life. Tree of eukaryotes showing major supergroups (indicated by color) together with graphics highlighting the morphological diversity of protists. Branches representing fungi, animals, and land plants are indicated by broken lines, with no further taxonomic resolution provided. The name charophytes was changed to streptophyte algae within the Archaeplastida. Adapted from [87], with permission of Yana Eglit and Patrick Keeling.

Primary photosynthetic eukaryotes (archaeplastids) acquired plastids from an ancient endosymbiosis with cyanobacteria and traditionally encompass glaucophyte algae, red algae, and the lineage of green algae and land plants [70,71]. Additional endosymbiotic events have spread red

algal-derived plastids via secondary or even higher-order endosymbioses among eukaryotes to disparate algal lineages. These include cryptophytes, haptophytes, alveolates, and stramenopiles [16,24], and lineages that still remain completely uncultivated, such as the 'deep-branching plastid lineages' [72], some of which are of high ecological importance or biotechnological relevance. Compared with heterotrophic eukaryotes, phototrophic species are generally easier to cultivate, given appropriate nutrients and light conditions. In fact, many are axenic. These cultures are also easier to maintain in the long term and their availability greatly facilitates laboratory procedures, giving access to ample fresh material for high-molecular-weight DNA and RNA extractions. Some of these organisms have fast division times and complete genomes are available, such as marine algal members of the Viridiplantae, which make them powerful models for studying features in land plants [73]. Sequencing of genomes from other diverse photosynthetic taxa will further our insights into photosynthetic processes, although in some cases will require overcoming challenges presented by thick cell walls with specialized compounds that reduce efficiency of nucleic acid extractions [74].

Snapshot of available protist genomes

One of the very first protist genomes to be sequenced was that of the apicomplexan parasite *Plasmodium*, published in 2002 [75]. The number of publicly accessible protist genomes has shown a steep increase since 2022 (Figure 1A, Key figure), especially within the oomycete plant pathogens and the diatoms (bacillariophytes), both belonging to the stramenopiles. Several initiatives have started to scale up the sequencing of protists (e.g., the Protist 10,000 Genome Project [76], the Aquatic Symbiosis Genomics Project [9], and the DToL framework). As of February 2025, Genomes on a Tree (GoaT; Box 3) lists 1 843 386 eukaryotic species, of which 72 599 belong to protists [77]. There are 46 184 eukaryotic assemblies corresponding to 20 004 species. Yet only 2743 of these assemblies correspond to 1121 protist species.

These initial resources reveal that the species-level diversity and complexity of protists extend to their genomes as well (Box 4), including a wide range of genome sizes. Released protist genomes show considerable variation in assembly spans (Figure 1B). Large genomes are especially prominent among dinoflagellates that possess a dinokaryon (Box 4). Most of the **primary assemblies** of protists belong to the SAR clade (stramenopiles, alveolates, rhizarians), with a majority of stramenopiles (mainly oomycetes, including the plant-pathogenic water mold *Phytophthora*) and alveolates (mostly apicomplexans like the parasitic *Plasmodium* and ciliates), as well as discobids (euglenozoans such as the parasitic *Leishmania*) genomes (Figure 1C). These genomes are often associated with research focused on crop and livestock/human diseases, which further compounds the relatively low number of rhizarian genome assemblies [78]. Most protist genomes are assembled at the contig or scaffold level, with only a few chromosome-level assemblies available, as, for example, for phototrophic species (chlorophytes and bacillariophytes) and parasitic groups (apicomplexans, euglenozoans, and oomycetes; Figure 1C). Additionally, Benchmarking Universal Single-Copy Orthologs (BUSCO) [79] completeness shows considerable variation (Figure 1D), with only about 2.7% of protist genomes achieving greater than 90% completion. However, the application of the BUSCO approach on protist genomes is expected to be limited intrinsically by the low number of protist genomic resources (see opportunities section for discussion). The release of long-read-based protist

Box 2. Practices with protist genomes

Different methodological approaches for protist genome acquisition exist, taking several aspects of the protists' life strategies and possible occurrence in cultures into account. Here, we provide a general graphical overview on used techniques (Figure I).

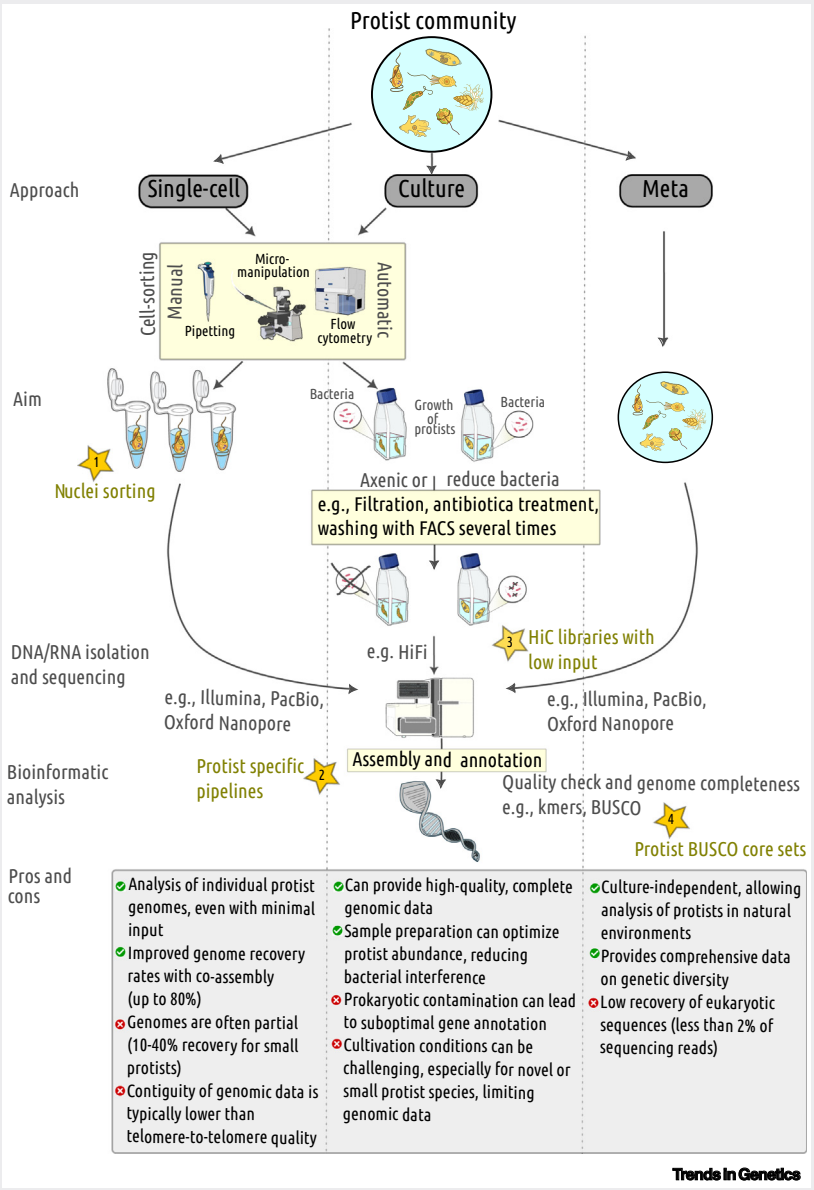
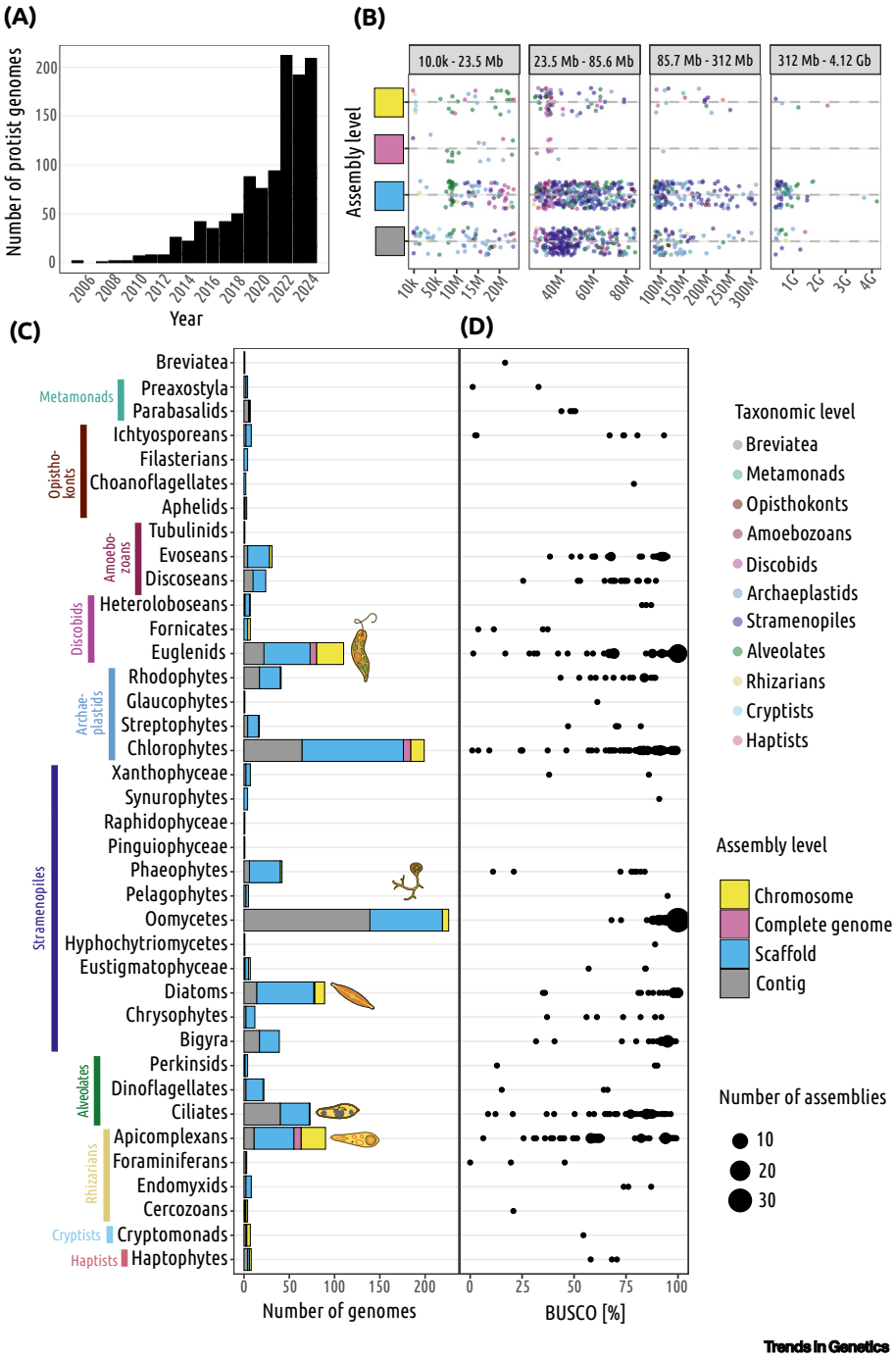


Figure I. Methodological approaches for protist genome acquisition. Current techniques for obtaining protist genomes from three starting points are illustrated here: single cells, cultures, and environmental samples (e.g., soil, water). Single-cell- and culture-based approaches require isolating cells through manual (e.g., micromanipulation) or higher throughput methods [e.g., fluorescence-activated cell sorting (FACS)], followed by amplification and sequencing of genomes [106] and transcriptomes. Cultivable protists can be grown in culture flasks by providing prey (e.g., bacteria) for heterotrophic protists. Afterwards bacteria must be reduced or axenic cultures created using methods such as filtration, antibiotic treatment, and repeated washing with FACS. DNA/RNA can be isolated from those cultures and sequenced. Metagenomics and metatranscriptomics allow culture-independent analysis, directly isolating DNA/RNA from environmental samples. Bioinformatic pipelines for assembly, annotation, and quality checks are necessary, with, for example, PacBio HiFi and Arima HiC data offering successful examples, though challenges in sequencing, assembly, and structural annotation persist. Yellow stars highlight long-term methodological development ideas discussed in section 4. Each of the three approaches has its own set of advantages and disadvantages, some of which are listed in the table below the figure. Protist illustrations were obtained from [87]; some illustrations (e.g., culture flasks, pipette) were adapted from NIAID NIH BIOART Source (bioart.niaid.nih.gov/bioart/).

Key figure

Available protist genomes from Genomes on a Tree (GoaT) retrieved November 2024)



(See figure legend at the bottom of the next page.)

Box 3. Databases for protist genomes

Protist genomes and related metadata can be accessed via several databases and browsers.

Genomes on a Tree (GoaT) (https://goat.genomehubs.org/) is a platform designed to provide genome-relevant metadata and sequencing project information for eukaryotic species. Developed to support the Earth BioGenome Project (EBP), GoaT aggregates validated metadata, including genome sizes, karyotypes, and sequencing status, from public sources and interpolates missing data using phylogenetic comparison [77]. GoaT offers a versatile API, web interface, and command line tools for data querying, exploration, and reporting, aiding large-scale genomic sequencing efforts and project coordination within the EBP framework.

The Protist 10,000 Genomes project (P10K) (https://ngdc.cncb.ac.cn/p10k/) aims to sequence genomes of 10 000 protist species [76]. P10K also provides a pipeline for single-cell genomics, including decontamination and annotation specifically for protists.

As part of the Ensembl database (developed by EMBL-EBI) EnsemblProtists (https://protists.ensembl.org/) is a specialized genome browser, which offers high-quality, annotated genomic and proteomic data for various protist species, sourced from the International Nucleotide Sequence Database Collaboration. It includes tools for visualizing genomic features and provides programmatic access via Perl and RESTful APIs.

The Eukaryotic Pathogen, Vector, and Host Informatics Resources (VEuPathDB) (https://veupathdb.org/veupathdb/app) is a comprehensive resource that provides access to genomic, transcriptomic, and proteomic data related to vector-borne and zoonotic pathogens [107].

EukProt (https://evocellbio.com/eukprot/) is a database of publicly available predicted protein sets selected from major eukaryotic supergroups (species are placed within the UniEuk taxonomic framework), designed to support gene-based research in areas like phylogenomics and gene family evolution [108].

genomes remains scarce, indicating that the protistology community is not yet fully benefiting from the significant sequencing technology advances of recent years.

Opportunities in the era of protist genomics

Current established standards for reference genomes target the highest assembly quality, with an expectation of chromosome-level resolution at maximal completeness for every species [1]. The focus of genomics techniques on metazoans and plants, and their respective molecular characteristics, is furthermore reflected in best practices and preferred software. Here we aim to challenge these current plant/animal-centered standards. When applying current pipelines to protist genomes, results often fall short of these standards, which speaks for the need of conceptual and methodological advancement. While no single approach is likely to generate high-quality genomes for the full spectrum of protist diversity, in this section we highlight opportunities for improvement and unique advances, both in wet-lab procedures and bioinformatic approaches.

Optimizing protist genome sequencing from cultures

Several methods help minimize contamination and maximize protist abundance in cultures, including antibiotic treatment (to obtain axenic or monoxenic cultures), size fractionation via filtration, and separation via FACS (Box 2). However, bacterial sequences often persist in non-axenic cultures, reducing sequencing depth and affecting genome assembly accuracy. Strategies such as targeting life-cycle stages with higher DNA/RNA yields and using

Figure 1. (A) Number of protist genomes generated over the years (from 2005 to 2024). (B) Assembly span of protist genomes categorized by assembly level (contig, scaffold, complete genome, chromosome), with colors representing different supergroups. Assembly spans are divided into different size ranges for clarity. (C) Barplot showing the number of protist genomes per taxonomic group within each supergroup, color-coded by assembly level. Protist illustrations were obtained from [87]. (D) Frequency distribution of Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness values (complete single-copies and duplicates) [%] for protist genome assemblies across taxonomic groups within the supergroups.

Box 4. Complexity of protist genomes

Protists often exhibit fascinating genomic features. Here, we provide three examples.

The kinetoplast, a DNA hyper-inflation in the mitochondrion

Kinetoplastids, a diverse group including obligatory parasitic trypanosomatids and mostly free-living bodonids, contains the agents of sleeping sickness (*Trypanosoma brucei* subspecies), Chagas disease (*Trypanosoma cruzi*), and leishmaniasis (*Leishmania* spp.) [109]. Kinetoplastids possess a kinetoplast, a massive body of mitochondrial DNA [110]. In trypanosomatids, the kinetoplast DNA includes maxicircles (20–25 kb, encoding mitoribosomal RNAs and oxidative phosphorylation proteins) and thousands of minicircles (0.5–10 kb) [111]. Diplonemids have large, circular, non-catenated mitochondrial DNA molecules, constituting up to 60% of their total cell DNA, representing the most DNA-rich organelles known [112]. Despite differences in DNA structure, both groups undergo extensive RNA editing – uridine insertions/deletions in kinetoplastids [113] and nucleotide substitutions in diplonemids [114] – indicating complex, potentially neutrally evolving mitochondrial machinery, unlike their Euglenida relatives, which have streamlined nucleic acids with standard coding capacity [115].

Nuclear dualism in ciliates

Ciliates have two types of nuclei, a small, diploid inactive germline micronucleus (MIC) and a large polyploid somatic macronucleus (MAC). The MIC is transcriptionally active during mating, participating in meiosis and sexual recombination, and passing on the genome during cell division. It also contains noncoding DNA **internally eliminated sequences (IESs)**, which are never expressed. The MAC serves as the active site for gene expression during vegetative growth, which is destroyed during sexual reproduction. During this process, some daughter MICs differentiate into MACs, eliminating IESs and undergoing chromosome fragmentation, rearrangement, and copy number amplification. In some ciliates, the macronuclear genome encodes 18 500 genes across 16 000 chromosomes [116,117].

The immense size of dinoflagellate genomes

Core dinoflagellates, related to ciliates and apicomplexans, have a unique nucleus called the dinokaryon with idiosyncratic genomic features not seen in other eukaryotes. Having abandoned histones for DNA packaging and using instead proteins of viral and bacterial origin [118,119], their genomes are tightly packed into liquid-crystalline chromosomes that are condensed throughout the whole cell cycle. Genome sizes range from 1000 to 215 000 Mb [120,121], with a portion of the genes arranged in tandem arrays [122]. Some of the smallest core dinoflagellate genomes known (1100–1500 Mb) belong to the ecologically important coral endosymbiont genus *Symbiodinium* [123]. Noncore dinoflagellates, like the parasite *Amoebophyra* spp. that infects various core dinoflagellates, have distinctly smaller genomes (~100 Mb) [124,125].

bioinformatic tools to remove contaminants, as demonstrated for the marine gregarine *Porospora gigantea* [58], can improve results.

When comprehensive genome sequencing is not yet feasible, transcriptomics combined with genome skimming, a cost-effective shallow sequencing approach, has the potential to provide valuable genetic insights, capturing high-copy regions, coding sequences, and conventional DNA barcodes for various applications.

Single-cell sequencing and metagenomics provide alternatives, especially for species that are not yet cultured, while improved protocols for high-molecular-weight DNA extraction (e.g., nuclei extraction) and high-quality Hi-C libraries with minimal starting material are urgently needed to enable chromatin studies from few or single cells.

Leveraging single-cell genomics for protists

Single-cell genomics enables the sequencing of rare and uncultured species. The feasibility of single-cell genomics with minimal input material was shown for four ciliate species where genomes were sequenced using two whole-genome amplification methods applied to individual cells [80]. Genome sequences of uncultured and dominant microeukaryotic species have also been obtained [57]. Though technically demanding, single-cell and pooled single-cell approaches enable precise isolation of individual protist cells using methods like micromanipulation, FACS, and laser-scanning technologies (e.g., confocal microscopy, laser capture microdissection; Box 2). When combined with single-cell techniques, these methods provide high-resolution

insights into gene expression, physiology, and evolution. While micromanipulation and FACS require time, expertise, and costly equipment, alternative approaches like selective lysis offer a promising, accessible way to enrich eukaryotic DNA by leveraging nuclear envelope protection. When combined with single-cell transcriptomics (sc RNA-seq), it helps mitigate challenges such as uneven genome amplification and better resolve repetitive regions. However, single-cell 'omics' data may still contain prokaryotic contamination, because of ingested, endosymbiotic, or surface-attached bacteria, and this signal will need to be removed bioinformatically.

The potential of metagenomics

Metagenome-assembled genomes (MAGs) have been successfully reconstructed from uncultured prokaryotic species, but only a few eukaryotic MAGs have been reported [81], often from low-diversity communities. Hundreds of planktonic marine eukaryote MAGs were recently assembled from a massive amount of sequencing data; these were combined with single-cell amplified genomes (SAGs) from dominant marine eukaryotes to generate a high-quality genomic repository [55]. While metagenomics pipelines were initially developed for large-scale sequencing of bacterial communities, specialized pipelines such as EukHeist [82] are being developed to better handle eukaryotic complexity, refine assembly approaches, and effectively target eukaryotic genomes. MAGs are so far commonly assembled from short-read metagenomes resulting in discontinuous, fragmented, contaminated, or unphased genome data, unable to resolve gene synteny, repetitive regions, polyploidy, large introns, etc. Metagenomic binning may perform poorly for some eukaryotic organisms and may fail to correctly bin some portions of the genome, including organellar genomes [83]. Long-read sequencing can overcome many of these challenges, providing a more complete and continuous picture of the metasample. Recent advancements in long-read sequencing, such as the PacBio Revio system and improved Oxford Nanopore chemistry, have made long-read metagenomics feasible.

Advancing assembly and annotation pipelines for protists

Assembly pipelines have advanced significantly in recent years, including their application to complex genomes, such as giant [84] or polyploid genomes [85]. Overall, available tools can produce high-quality assemblies when provided with high-quality genome data. In this regard, chromatin-capture techniques play a crucial role in improving assembly quality and urgently require adapted low-input protocols (see earlier). Also, data decontamination is essential when genomic data have been collected from symbiotic species (see earlier). Software like Tiara [86] can help reduce bacterial contamination, but decontamination pipelines must also account for and specifically detect horizontal and endosymbiotic gene transfer between protists and bacteria. Here, we advocate for the development of protist group-specific data decontamination tools.

Genome assembly quality is typically assessed using the eukaryotic BUSCO core set [79]. However, few core sets are specifically designed for protist groups (e.g., stramenopiles, euglenozoans, apicomplexans, chlorophytes, and alveolates), limiting the accuracy of these evaluations and resulting in underestimates of completeness. This issue stems from the scarcity and bias of available protist genomes, as well as their high genetic divergence. Expanding and refining taxon-specific BUSCO core gene sets is essential for improving the evaluation of assembly completeness.

Annotation pipelines are typically designed for animal and plant model organisms, reflecting their gene architectures. As a result, they may not align with the unique gene features of protists (Box 4), leading to lower-quality gene model predictions. Additionally, many protist genes remain unclassified due to limited genomic and transcriptomic databases, scarce reference genomes, and their distinct gene architectures.

Expanding genomic and transcriptomic resources in databases is a crucial prerequisite for high-quality genome assemblies and annotations in the long term. We advocate for publication of fragmented and/or incomplete transcriptome and genome assemblies – sub-optimal as they are – in order to support the development of databases. This intermediate step is essential for advancing methodological development, as incomplete databases are one major obstacle in achieving the highest quality annotations in the future. The availability of lower-quality resources should be seen as a necessary intermediate step toward eventual improvement.

Concluding remarks and future perspectives

Advancing biodiversity genomics in general, and protist genomics in particular, will require global collaboration under the EBP and ongoing support from funding agencies and philanthropists. While molecular biologists and bioinformaticians are key, taxonomists will play a crucial role for taxonomic validation. The unique ecological, molecular, and genomic diversity of protists offers significant potential to address major questions in ecology and evolution. This diversity requires careful consideration to advance biodiversity genomics, presenting an opportunity to develop theory, technology, and bioinformatics that can accommodate protist genomes, which are challenging to sequence and assemble (see Outstanding questions).

Generating high-quality genomes from culturable protists is important, but since most protists cannot be cultured, efforts should prioritize improving methods to handle small biological samples to obtain single-species and high-quality genome data. In parallel, we advocate for optimizing single-cell genomics and transcriptomics for the application with protists. The genomes generated using current single-cell techniques may not meet the strict criteria for reference genomes as defined by EBP, but they will serve as valuable referential genomes for understanding the diversity of protist genomes. Furthermore, we strongly advocate for the release of incomplete genome and transcriptome assemblies from a broader protist taxon spectrum to address database limitations, which are a major barrier to achieving high-quality genome annotations. This can also be supported by leveraging higher quality MAGs via long-read metagenomics. The required development of taxon-specific BUSCO core gene sets will only then become possible to allow for the accurate evaluation of completeness in ultimately high-quality protist genomes.

Increasing protist genomic resources by adopting the here-addressed recommendations is key to advancing our understanding of protist diversity and evolution. Ultimately, achieving the highest quality genome assemblies for protists must remain the goal for their integration into the global biodiversity genomics framework.

Outstanding questions

How can genomic resources for protists be utilized to decode ecosystem dynamics, particularly through metagenomics and metatranscriptomics, to better understand their roles in biogeochemical cycles and microbial food webs?

How much gene transfer, both from the endosymbiont (EGT) and from other sources (LGT), took place in different eukaryotic lineages?

In what ways can the expanding knowledge of noncanonical genetic codes and unique molecular features in protists contribute to the development of new genetic tools and biotechnologies?

What are the common and lineage-specific genomic adaptations across different eukaryotic groups that are involved in the terrestrialization of life?

How can we adapt chromatin-capture techniques for ultra-low input or even single-cell dimensions for protists?

How can bioinformatic pipelines be standardized to handle incomplete gene annotations while remaining flexible to the molecular diversity of protists, including newly discovered taxa?

Is the current EBP definition of a reference genome too strict, and would broadening it accelerate biodiversity genomics for challenging organism groups?

Declaration of interests

The authors declare no competing interests.

References

1. Theissinger, K. *et al.* (2023) How genomics can help biodiversity conservation. *Trends Genet.* 39, 545–559
2. Blaxter, M. *et al.* (2022) Why sequence all eukaryotes? *Proc. Natl. Acad. Sci. U. S. A.* 119, e2115636118
3. Lewin, H.A. *et al.* (2022) The Earth BioGenome Project 2020: starting the clock. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2115635118
4. Formenti, G. *et al.* (2022) The era of reference genomes in conservation genomics. *Trends Ecol. Evol.* 37, 197–202
5. Guiglielmoni, N. *et al.* (2022) A deep dive into genome assemblies of non-vertebrate animals. *Peer Commun. J.* 2, e29
6. Rhie, A. *et al.* (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737–746
7. Worden, A.Z. *et al.* (2015) Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science* 347, 1257594
8. Bachy, C. *et al.* (2022) Marine protists: a hitchhiker's guide to their role in the marine microbiome. In *The Marine Microbiome* (3) (Stal, L.J. and Cretoiu, M.S., eds), pp. 159–241, Springer International Publishing
9. McKenna, V. *et al.* (2024) The Aquatic Symbiosis Genomics Project: probing the evolution of symbiosis across the Tree of Life [version 2; peer review: 1 approved, 1 approved with reservations]. *Wellcome Open Res.* 6, 254
10. del Campo, J. *et al.* (2014) The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.* 29, 252–259

11. Massana, R. *et al.* (2022) Protists, la principal font de diversitat genòmica en eucariotes. *Treb. Soc. Catalana Biol.* 72, 43–50
12. Sibbald, S.J. and Archibald, J.M. (2017) More protist genomes needed. *Nat. Ecol. Evol.* 1, 0145
13. Adl, S.M. *et al.* (2018) Revisions to the classification, nomenclature, and diversity of eukaryotes. *J. Eukaryot. Microbiol.* 66, 4–119
14. Hofstetter, V. *et al.* (2007) Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: a case study of the Lecanoromycetes (Ascomycota). *Mol. Phylogenet. Evol.* 44, 412–426
15. Rajendhran, J. and Gunasekaran, P. (2011) Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. *Microbiol. Res.* 166, 99–110
16. Strassert, J.F.H. *et al.* (2021) A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* 12, 1879
17. Ocaña-Pallarès, E. *et al.* (2022) Divergent genomic trajectories predate the origin of animals and fungi. *Nature* 609, 747–753
18. Tovar, J. *et al.* (2003) Mitochondrial remnant organelles of *Giardia* function in iron-sulphur protein maturation. *Nature* 426, 172–176
19. Williamson, K. *et al.* (2025) A robustly rooted tree of eukaryotes reveals their excavate ancestry. *Nature* 640, 974–981
20. Bartošová-Sojková, P. *et al.* (2024) Inside the host: understanding the evolutionary trajectories of intracellular parasitism. *Ann. Rev. Microbiol.* 78, 39–59
21. Jackson, A.P. *et al.* (2016) Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. *Curr. Biol.* 26, 161–172
22. Nowack, E.C.M. and Weber, A.P.M. (2018) Genomics-informed insights into endosymbiotic organelle evolution in photosynthetic eukaryotes. *Annu. Rev. Plant Biol.* 69, 51–84
23. Archibald, J.M. (2015) Endosymbiosis and eukaryotic cell evolution. *Curr. Biol.* 25, R911–R921
24. Sibbald, S.J. and Archibald, J.M. (2020) Genomic insights into plastid evolution. *Genome Biol. Evol.* 12, 978–990
25. Richards, T.A. *et al.* (2024) Reconstructing the last common ancestor of all eukaryotes. *PLoS Biol.* 22, e3002917
26. Donoghue, P.C.J. *et al.* (2023) Defining eukaryotes to dissect eukaryogenesis. *Curr. Biol.* 33, R919–R929
27. Leger, M.M. and Ruiz-Trillo, I. (2022) Phylogenetics of clonal multicellularity. In *The Evolution of Multicellularity* (1st edn) (Herron, M.D. *et al.*, eds), pp. 157–186, CRC Press
28. Lamża, Ł. (2023) Diversity of 'simple' multicellular eukaryotes: 45 independent cases and six types of multicellularity. *Biol. Rev.* 98, 2188–2209
29. Torruella, G. *et al.* (2015) Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Curr. Biol.* 25, 2404–2410
30. Carr, M. *et al.* (2017) A six-gene phylogeny provides new insights into choanoflagellate evolution. *Mol. Phylogenet. Evol.* 107, 166–178
31. Bringloe, T.T. *et al.* (2020) Phylogeny and evolution of the brown algae. *Crit. Rev. Plant Sci.* 39, 281–321
32. Hess, S. *et al.* (2022) A phylogenomically informed five-order system for the closest relatives of land plants. *Curr. Biol.* 32, 4473–4482.e7
33. Bierenbroodspot, M.J. *et al.* (2024) Phylogeny and evolution of streptophyte algae. *Ann. Bot.* 134, 385–400
34. Nagy, L.G. *et al.* (2018) Complex multicellularity in fungi: evolutionary convergence, single origin, or both? *Biol. Rev.* 93, 1778–1794
35. Niklas, K.J. and Newman, S.A. (2020) The many roads to and from multicellularity. *J. Exp. Bot.* 71, 3247–3253
36. Batista, R.A. *et al.* (2024) Insights into the molecular bases of multicellular development from brown algae. *Development* 151, dev203004
37. Adamska, M. *et al.* (2007) The evolutionary origin of hedgehog proteins. *Curr. Biol.* 17, R836–R837
38. Grau-Bové, X. *et al.* (2017) Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* 6, e26036
39. López-Escardó, D. *et al.* (2019) Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. *Philos. Trans. R. Soc. B Biol. Sci.* 374, 20190088
40. Culbertson, E.M. and Levin, T.C. (2023) Eukaryotic CD-NTase, STING, and viperin proteins evolved via domain shuffling, horizontal transfer, and ancient inheritance from prokaryotes. *PLoS Biol.* 21, e3002436
41. Nedelcu, A.M. (2019) Independent evolution of complex development in animals and plants: deep homology and lateral gene transfer. *Dev. Genes Evol.* 229, 25–34
42. Suga, H. *et al.* (2013) The *Capsaspora* genome reveals a complex unicellular prehistory of animals. *Nat. Commun.* 4, 2325
43. Dadras, A. *et al.* (2023) Environmental gradients reveal stress hubs pre-dating plant terrestrialization. *Nat. Plants* 9, 1419–1438
44. Feng, X. *et al.* (2024) Genomes of multicellular algal sisters to land plants illuminate signaling network evolution. *Nat. Genet.* 56, 1018–1031
45. King, N. *et al.* (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451, 783–788
46. Richter, D.J. *et al.* (2018) Gene family innovation, conservation and loss on the animal stem lineage. *eLife* 7, e34226
47. Jamy, M. *et al.* (2022) Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Nat. Ecol. Evol.* 6, 1458–1470
48. Naranjo-Ortiz, M.A. and Gabaldón, T. (2019) Fungal evolution: major ecological adaptations and evolutionary transitions. *Biol. Rev.* 94, 1443–1476
49. Záhonová, K. *et al.* (2025) Comparative genomic analysis of trypanosomatid protists illuminates an extensive change in the nuclear genetic code. *mBio*, e00885-25
50. Faktorová, D. *et al.* (2020) Genetic tool development in marine protists: emerging model organisms for experimental cell biology. *Nat. Methods* 17, 481–494
51. Lukeš, J. *et al.* (2023) Trypanosomes as a magnifying glass for cell and molecular biology. *Trends Parasitol.* 39, 902–912
52. Clayton, C. (2019) Regulation of gene expression in trypanosomatids: living with polycistronic transcription. *Open Biol.* 9, 190072
53. Coale, T.H. *et al.* (2024) Nitrogen-fixing organelle in a marine alga. *Science* 384, 217–222
54. Carradec, Q. *et al.* (2018) A global ocean atlas of eukaryotic genes. *Nat. Commun.* 9, 373
55. Delmont, T.O. *et al.* (2022) Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics* 2, 100123
56. Merz, E. *et al.* (2023) Disruption of ecological networks in lakes by climate change and nutrient fluctuations. *Nat. Clim. Chang.* 13, 389–396
57. Labarre, A. *et al.* (2021) Comparative genomics reveals new functional insights in uncultured MAST species. *ISME J.* 15, 1767–1781
58. Boisard, J. *et al.* (2022) Marine gregarine genomes reveal the breadth of apicomplexan diversity with a partially conserved glideosome machinery. *BMC Genomics* 23, 485
59. Thomé, P.C. *et al.* (2024) Phylogenomics including new sequence data of phytoplankton-infecting chytrids reveals multiple independent lifestyle transitions across the phylum. *Mol. Phylogenet. Evol.* 197, 108103
60. Mongue, A.J. *et al.* (2023) Genome sequence of *Ophryocystis elektroscirrha*, an apicomplexan parasite of monarch butterflies: cryptic diversity and response to host-sequestered plant chemicals. *BMC Genomics* 24, 278
61. Templeton, T.J. *et al.* (2010) A genome-sequence survey for *Ascogregarina taiwanensis* supports evolutionary affiliation but metabolic diversity between a gregarine and *Cryptosporidium*. *Mol. Biol. Evol.* 27, 235–248
62. Nkhoma, S.C. *et al.* (2020) Co-transmission of related malaria parasite lineages shapes within-host parasite diversity. *Cell Host Microbe* 27, 93–103.e4
63. Negreira, G.H. *et al.* (2022) High throughput single-cell genome sequencing gives insights into the generation and evolution of mosaic aneuploidy in *Leishmania donovani*. *Nucleic Acids Res.* 50, 293–305
64. Wood, D.E. *et al.* (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257

65. Jousset, A. *et al.* (2006) Secondary metabolites help biocontrol strain *Pseudomonas fluorescens* CHA0 to escape protozoan grazing. *Appl. Environ. Microbiol.* 72, 7083–7090
66. Glücksman, E. *et al.* (2010) Closely related protist strains have different grazing impacts on natural bacterial communities. *Environ. Microbiol.* 12, 3105–3113
67. Hori, R.S. *et al.* (2021) Growth pattern of the siliceous skeletons of living Spumellaria (Radiolaria) from the Kuroshio Current, offshore southwestern Shikoku Island, Japan. *Rev. Micropaleontol.* 71, 100504
68. Sykes, F.E. *et al.* (2024) Large-scale culturing of the subpolar foraminifera *Globigerina bulloides* reveals tolerance to a large range of environmental parameters associated to different life-strategies and an extended lifespan. *J. Plankton Res.* 46, 403–420
69. Decelle, J. *et al.* (2012) An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U. S. A.* 109, 18000–18005
70. Irisarri, I. *et al.* (2021) Phylogenomic insights into the origin of primary plastids. *Syst. Biol.* 71, 105–120
71. Schön, M.E. *et al.* (2021) Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nat. Commun.* 12, 6651
72. Choi, C.J. *et al.* (2017) Newly discovered deep-branching marine plastid lineages are numerically rare but globally distributed. *Curr. Biol.* 27, R15–R16
73. Bachy, C. *et al.* (2022) The land–sea connection: insights into the plant lineage from a green algal perspective. *Annu. Rev. Plant Biol.* 73, 585–616
74. Moore, M. and Malia Moore, T.S.S. (2023) High molecular weight DNA extraction for marine macroalgal tissue. *protocols.io*, Published online May 8, 2023. https://doi.org/10.17504/protocols.io.14egn2dnpg5d/v1
75. Gardner, M.J. *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419, 498–511
76. Gao, X. *et al.* (2024) The P10K database: a data portal for the protist 10 000 genomes project. *Nucleic Acids Res.* 52, D747–D755
77. Challis, R. *et al.* (2023) Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic tree of life. *Wellcome Open Res.* 8, 24
78. Burki, F. and Keeling, P.J. (2014) Rhizaria. *Curr. Biol.* 24, R103–R107
79. Manni, M. *et al.* (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* 38, 4647–4654
80. Chen, W. *et al.* (2021) The hidden genomic diversity of ciliated protists revealed by single-cell genome sequencing. *BMC Biol.* 19, 264
81. Massana, R. and López-Escardó, D. (2022) Metagenome assembled genomes are for eukaryotes too. *Cell Genomics* 2, 100130
82. Alexander, H. *et al.* (2023) Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *mBio* 14, e01676-23
83. Jamy, M. *et al.* (2025) New deep-branching environmental plastid genomes on the algal tree of life. *bioRxiv*, Published online January 18, 2025. https://doi.org/10.1101/2025.01.16.633336
84. Schartl, M. *et al.* (2024) The genomes of all lungfish inform on genome expansion and tetrapod evolution. *Nature* 634, 96–103
85. Serra Mari, R. *et al.* (2024) Haplotype-resolved assembly of a tetraploid potato genome using long reads and low-depth offspring data. *Genome Biol.* 25, 26
86. Karlicki, M. *et al.* (2022) Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* 38, 344–350
87. Keeling, P.J. and Eglit, Y. (2023) Openly available illustrations as tools to describe eukaryotic microbial diversity. *PLoS Biol.* 21, e3002395
88. Goldfuß, G.A. (1818) Ueber die classification der zoophyten. *Isis Oder Encycl. Ztg. Von Oken* 2, 1008–1013
89. Goldfuß, A. (1820) *Handbuch der Zoologie: Erste Abteilung*, Johann Leonhard Schrag
90. Haeckel, E. (1866) *Generelle Morphologie der Organismen*, de Gruyter
91. Whittaker, R.H. (1969) New concepts of kingdoms of organisms: evolutionary relations are better represented by new classifications than by the traditional two kingdoms. *Science* 163, 150–160
92. Archibald, J.M. *et al.*, eds (2017) *Handbook of the Protists*, 2nd edn Springer
93. Burki, F. *et al.* (2020) The new tree of eukaryotes. *Trends Ecol. Evol.* 35, 43–55
94. Adl, S.M. *et al.* (2007) Diversity, nomenclature, and taxonomy of protists. *Syst. Biol.* 56, 684–689
95. de Vargas, C. *et al.* (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* 348, 1261605
96. Grossmann, L. *et al.* (2016) Protistan community analysis: key findings of a large-scale molecular sampling. *ISME J.* 10, 2269–2279
97. Mahé, F. *et al.* (2017) Parasites dominate hyperdiverse soil protist communities in neotropical rainforests. *Nat. Ecol. Evol.* 1, 0091
98. Massana, R. *et al.* (2014) Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.* 8, 854–866
99. Gawryluk, R.M.R. *et al.* (2016) Morphological identification and single-cell genomics of marine diplonemids. *Curr. Biol.* 26, 3053–3059
100. Singer, D. *et al.* (2021) Protist taxonomic and functional diversity in soil, freshwater and marine ecosystems. *Environ. Int.* 146, 106262
101. Garner, R.E. *et al.* (2022) Protist diversity and metabolic strategy in freshwater lakes are shaped by trophic state and watershed land use on a continental scale. *mSystems* 7, e00316-22
102. Cordier, T. *et al.* (2022) Patterns of eukaryotic diversity from the surface to the deep-ocean sediment. *Sci. Adv.* 8, eabj9309
103. Schoenle, A. *et al.* (2021) High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. *Commun. Biol.* 4, 501
104. Acosta, E. *et al.* (2024) Protist diversity and co-occurrence patterns obtained by metabarcoding of terricolous lichens, coastal cliffs and a microbial mat in the Atacama Desert, northern Chile. *Eur. J. Protistol.* 95, 126108
105. Campo, J. del *et al.* (2023) The protist cultural renaissance. *Trends Microbiol.* 32, 128–131
106. Sieracki, M.E. *et al.* (2019) Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci. Rep.* 9, 6025
107. Harb, O.S. *et al.* (2024) VEuPathDB resources: a platform for free online data exploration, integration, and analysis. In *Comparative Genomics: Methods and Protocols* (Setubal, J.C. *et al.*, eds), pp. 573–586, Springer
108 Richter, D.J. *et al.* (2022) EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community J.* 2, e56
109. Flegontova, O. *et al.* (2020) Environmental determinants of the distribution of planktonic diplonemids and kinetoplastids in the oceans. *Environ. Microbiol.* 22, 4014–4031
110. Jensen, R.E. and Englund, P.T. (2012) Network news: the replication of kinetoplast DNA. *Ann. Rev. Microbiol.* 66, 473–491
111. Butenko, A. *et al.* (2024) Mitochondrial genomes revisited: why do different lineages retain different genes? *BMC Biol.* 22, 15
112. Lukeš, J. *et al.* (2018) Massive mitochondrial DNA content in diplonemid and kinetoplastid protists. *IUBMB Life* 70, 1267–1274
113. Aphasizheva, I. *et al.* (2020) Lexis and grammar of mitochondrial RNA processing in trypanosomes. *Trends Parasitol.* 36, 337–355
114. Kaur, B. *et al.* (2020) Gene fragmentation and RNA editing without borders: eccentric mitochondrial genomes of diplonemids. *Nucleic Acids Res.* 48, 2694–2708

115. Dobáková, E. *et al.* (2015) Unexpectedly streamlined mitochondrial genome of the euglenozoan *Euglena gracilis*. *Genome Biol. Evol.* 7, 3358–3367
116. Boscaro, V. and Keeling, P.J. (2023) How ciliates got their nuclei. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2221818120
117. Seah, B.K.B. *et al.* (2023) MITE infestation accommodated by genome editing in the germline genome of the ciliate *Blepharisma*. *Proc. Natl. Acad. Sci. USA* 120, e2213985120
118. Gornik, S.G. *et al.* (2012) Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. *Curr. Biol.* 22, 2303–2312
119. Wong, J.T.Y. *et al.* (2003) Histone-like proteins of the dinoflagellate *Crypthecodinium cohnii* have homologies to bacterial DNA-binding proteins. *Eukaryot. Cell* 2, 646–650
120. Lin, S. (2024) A decade of dinoflagellate genomics illuminating an enigmatic eukaryote cell. *BMC Genomics* 25, 932
121. Roy, S.W. *et al.* (2023) Intron-rich dinoflagellate genomes driven by Introner transposable elements of unprecedented diversity. *Curr. Biol.* 33, 189–196.e4
122. Beauchemin, M. *et al.* (2012) Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic. *Proc. Natl. Acad. Sci. USA* 109, 15793–15798
123. Liu, H. *et al.* (2018) *Symbiodinium* genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Commun. Biol.* 1, 95
124. Farhat, S. *et al.* (2021) Rapid protein evolution, organellar reductions, and invasive intronic elements in the marine aerobic parasite dinoflagellate *Amoebophrya* spp. *BMC Biol.* 19, 1
125. John, U. *et al.* (2019) An aerobic eukaryotic parasite with functional mitochondria that likely lacks a mitochondrial genome. *Sci. Adv.* 5, eaav1110