The New Tree of Eukaryotes

Fabien Burki,1,2,4,* Andrew J. Roger,3,5 Matthew W. Brown,5,6 and Alastair G.B. Simpson4,7,*

For 15 years, the eukaryote Tree of Life (eToL) has been divided into five to eight major groupings, known as ‘supergroups’. However, the tree has been profoundly rearranged during this time. The new eToL results from the widespread application of phylogenomics and numerous discoveries of major lineages of eukaryotes, mostly free-living heterotrophic protists. The evidence that supports the tree has transitioned from a synthesis of molecular phylogenetics and biological characters to purely molecular phylogenetics. Most current supergroups lack defining morphological or cell-biological characteristics, making the supergroup label even more arbitrary than before. Going forward, the combination of traditional culturing with maturing culture-free approaches and phylogenomics should accelerate the process of completing and resolving the eToL at its deepest levels.

The Eukaryote Tree of Life

Resolving the evolutionary tree for all eukaryotes has been a long-standing goal in biology. Inferring an eToL that is both accurate and comprehensive is a worthwhile objective in itself, but the eToL is also the framework on which we understand the origins and history of eukaryote biology and the evolutionary processes underpinning it. It is therefore a fundamental tool for studying many aspects of eukaryote evolution, such as cell biology, genome organization, sex, and multicellularity. In the molecular era, the eToL has also become a vital resource to interpret environmental sequence data and thus reveal the diversity and composition of ecological communities.

Although most of the described species of eukaryotes belong to the multicellular groups of animals (Metazoa), land plants, and fungi, it has long been clear that these three ‘kingdoms’ represent only a small proportion of high-level eukaryote diversity. The vast bulk of this diversity – including dozens of extant ‘kingdom-level’ taxa – is found within the ‘protists’, the eukaryotes that are not animals, plants, or fungi [1–6]. To a first approximation, inferring the eToL is to resolve the relationships among the major protist lineages. However, this task is complicated by the fact that protists are much less studied overall than animals, plants, or fungi [7]. Molecular sequence data has accumulated slowly for many known protist taxa and numerous important lineages were completely unknown (or were not cultivated, hence challenging to study) when the molecular era began. Thus, resolving the eToL has been a process where large-scale discovery of major lineages has occurred simultaneously with deep-level phylogenetic inference. This makes the task at hand analogous to a jigsaw puzzle, but one where a large and unknown number of pieces are missing from the box and instead are hidden under various pieces of the furniture.

The Supergroups Model

By the early 2000s, a model of the tree emerged that divided almost all of known eukaryote diversity among five to eight major taxa usually referred to as ‘supergroups’ [8–12]. The category of supergroup was a purely informal one, denoting extremely broad assemblages that contain, for example, the traditional ‘kingdoms’ like Metazoa and Fungi as subclades. Thus, the original supergroups generally represented the most inclusive collections of organisms within eukaryotes for which there was reasonable evidence that they formed a monophyletic group. A typical list of these groups included (with some differences in capitalization and endings): Archaeplastida (also known as Plantae), Chromalveolata, Rhizaria (or Cercozoa), Opisthokonta, Amoebozoa, and Excavata (see Box 1 for short descriptions). The main variations between accounts from that time were that some united Opisthokonta and Amoebozoa as ‘unikonts’ [12] (much later renamed ‘Amorpha’ [13]) or did not show Excavata and/or Chromalveolata confidently resolved as clades [10,11]. For half of the groups (i.e., Opisthokonta, Amoebozoa, and Rhizaria), the principal evidence supporting their unity was the phylogenies of one or a few genes [14–16]. For the others, it was a combination of

Highlights

The eukaryote Tree of Life (eToL) represents the phylogeny of all eukaryotic lineages, with the vast bulk of this diversity comprising microbial ‘protists’. Since the early 2000s, the eToL has been summarized in a few (five to eight) ‘supergroups’. Recently, this tree has been deeply remodeled due mainly to the maturation of phylogenomics and the addition of numerous new ‘kingdom-level’ lineages of heterotrophic protists.

The current eToL is derived almost exclusively from molecular phylogenies, in contrast to earlier models that were synthesis of molecular and other biological data.

The supergroup model for the eToL has become increasingly abstract due to the absence of known shared derived characteristics for the new supergroups.

Culture-based studies, not higher-throughput methods, have been responsible for most of the new major lineages recently added to the eToL.

1Department of Organismic Biology, Program in Systematic Biology, Uppsala University, Uppsala, Sweden
2Science for Life Laboratory, Uppsala University, Uppsala, Sweden
3Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS, Canada
4Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS, Canada
5Department of Biological Sciences, Mississippi State University, Mississippi State, MS, USA
6Institute for Genomics, Biocomputing, and Biotechnology, Mississippi State University, Mississippi State, MS, USA
7Department of Biology, Dalhousie University, Halifax, NS, Canada
*Correspondence: fabien.burki@ebc.uu.se, alastair.simpson@dal.ca

@Twitter: @fburki (F. Burki).
weaker molecular phylogenetic evidence and shared derived cell-biological features. Archaeplas-
tida and Chromalveolata were each identified by the presence of similar plastids [17,18], with se-
quences from plastid genomes supporting an ancestral endosymbiotic origin of plastids in each
group [19,20]. Excavata, meanwhile, was distinguished by the inference that taxa shared a derived,
complex flagellar apparatus cytoskeleton [21]. Consequently, the original supergroup-based eToLs
were syntheses of different information rather than straightforward summaries of molecular
phylogenies.

The supergroup model for the eToL became widely popular in both the primary literature and text-
books, for several reasons. First, the model made for convenient and efficient summaries of eu-
karyotes, since almost all species fell into these few relatively diverse major groups. Second, all of
the original supergroups, except Rhizaria, had at least one distinctive biological characteristic
that seemed to ancestrally define them (see above and Box 1). Third, the groupings seemed to coincide
with the limits of phylogenetic resolution. In fact, the overarching supergroup model has remained
the standard description of the eToL for 15 years, despite major changes in our knowledge of eu-
karyotic phylogeny and diversity over that time.

Box 1. The Original Supergroups – and Where Are They Now?
Five to six supergroups were originally proposed, depending on whether Opisthokonta and Amoebozoa were
unified in the larger group unikonts [9,12]. The name unikonts (based on a now-discarded hypothesis of a uni-
flagellated ancestor) was later replaced by Amorphea [13]. The six supergroups version corresponded to the
following.

- Opisthokonta includes animals, fungi, and several protist lineages that are most closely related to either
animals or fungi. Opisthokonta remains a robust clade in modern phylogenies; however, it is nested within
at least two larger taxa, Amorphea and Obazoa, that are frequently treated as supergroups instead.

- Amoebozoa is also still a robust group, but now is often regarded as a member of the supergroup Amor-
phea. Amoebozoa includes free-living amoeboid forms with lobose pseudopodia (e.g., Amoeba) but
also more filose amoebae, some flagellates, and various slime molds.

- Excavata was originally proposed based on a distinctive morphology, namely a particular feeding groove
form and associated cytoskeleton system, found in many enigmatic flagellated protists. Phylogenetics
and phylogenomics defined three monophyletic subgroups – Discoba, Metamonada, and malawimonads
– but have not consistently placed them together as a single clade. The name is now usually restricted to
a Discoba–Metamonada clade (quite possibly artificial; see main text) or regarded as referring to a paraphy-
letic group.

- Archaeplastida are distinguished by the presence of primary plastids – the photosynthetic organelles
deriving directly from cyanobacteria by endosymbiosis. The three main groups with primary plastids are
the green algae and land plants, red algae (and likely their recently discovered relative Rhodelphis), and
glaucophyte algae. Today, Archaeplastida is generally still considered a supergroup, although most phy-
logenomic analyses do not strongly support its monophyly (i.e., all three host lineages forming a single
clade to the exclusion of other supergroups).

- Chromalveolata contained groups with red alga-derived secondary plastids (i.e., Alveolata, Stramenopila,
Haptophyta, and Cryptophyta). This group was based on the assumption that these plastids were acquired
once in a common ancestor, which was supported by plastid evidence but never strongly from the host
perspective. Chromalveolata has been shown to be polyphyletic, with Alveolata and Stramenopila
belonging to Sar (in TSAR), Haptophyta in Haptista, and Cryptophyta in Cryptista.

- Rhizaria was the latest addition at the time the supergroup model was proposed. It includes a wide diversity
of amoebae (e.g., foraminiferans, the radiolarians, filose testate amoebae), flagellates, various parasites,
and the chlorarachniophyte algae. In contrast to all other original supergroups, which were at least partly
distinguished by morphological characters, Rhizaria was inferred more or less exclusively using molecular
phylogenetics. It is now part of Sar (in TSAR) along with Alveolata and Stramenopila.
New Methods and Taxa

The profound changes to the eToL have come from the development of phylogenomics and, concomitantly, the addition of many evolutionarily important protist lineages into molecular datasets. Below, we briefly introduce these two aspects.

Phylogenomics

The term ‘phylogenomics’ covers various approaches combining genomic-scale data with phylogenetic methods. In the context of the eToL, it usually refers to the estimation of organismal phylogeny from datasets containing dozens to hundreds of gene alignments, most often nucleus-encoded genes analyzed as inferred amino acid sequences [22]. The data are sourced from a mixture of genome and, frequently, transcriptome sequencing projects. The introduction of phylogenomics offered the promise of overcoming the limited information afforded by single genes, which were mostly inadequate to resolve deep divergences within the eToL [23]. However, voices warned early on that most of the analysis artefacts known to afflict single-gene phylogenies can also apply to phylogenomics [24]. Phenomena that cause unrelated taxa to cluster together in phylogenies, such as compositional bias and high rates of sequence divergence, often also affect the whole genome. Therefore, merely adding genes can amplify artefacts rather than overriding them [25]. Accuracy might be improved by using more realistic evolutionary models, and especially by careful choice of taxa, where this is possible (see below). Examining multiple genes also raises the specter of combining different gene histories together artificially, making careful quality controls essential to eliminate incorrect paralog assignments, contaminating sequences, etc. (see Box 2 for a typical ‘phylogenomic pipeline’).

Box 2. Example of a Phylogenomic Analysis Pipeline

Construction of datasets for phylogenomics is complicated, requiring painstaking care to exclude spurious data (e.g., contaminants, paralogs) and select taxa appropriately. Deep-level phylogenomic analyses typically use inferred amino acid sequences of proteins, and sets of hundreds of widely present and/or highly expressed proteins are curated by various research groups. When new taxa are added, homologous sequences are retrieved from their transcriptomic or predicted gene sets, usually using pairwise alignment similarity tools (e.g., BlastP, Diamond-BlastP) or profile-based approaches (e.g., hidden Markov model methods like HMM-search). Typically, a series of checks are made to exclude paralogous sequences, often through reciprocal best BlastP hit to a set of manually curated orthologs. The proteins from new taxa that pass these checks are provisionally considered orthologous and are aligned with those from the hundreds of species in the curated dataset. Maximum likelihood (ML) trees are then estimated for each gene alignment, with bootstrapping to assess branch support. These trees are examined to identify and exclude sequences with apparent or actual evolutionary histories that differ from the organismal phylogeny, such as lateral gene transfers, incorrect paralog selections, and various contaminants. Contamination may occur during sequencing (referred to as on-sequencer/flow cell contamination), during library preparation, or in cell culture. These gene tree examinations currently include laborious by-eye inspections of the phylogenies, since some aspects still require human interpretation and decisions. A suitable subset of taxa is then selected for the actual analysis. This selection aims to evenly cover the relevant phylogenetic breadth while excluding problematic species (e.g., those with limited data, extreme evolutionary rates in many genes, etc.). The explosion in the number of species for which omic data are available has greatly enhanced choice in taxon selection, as well as the detection (and elimination) of nonvertical signals in the data (e.g., [54,56,62]).

Dataset assembly is followed by the actual phylogenomic analyses, in which hundreds of genes are concatenated into a phylogenomic ‘supermatrix’. Usually, both ML and Bayesian analyses are conducted. Various evolutionary models are employed, with choice often constrained by computational logistics. Site-heterogeneous models, in which the profile of substitution propensities can differ among sites in the alignment, appear to be particularly important for improved phylogenetic accuracy. These models were first implemented in the Bayesian inference platform PhyloBayes [102], but the analyses are computationally intensive and problems with mixing and convergence are common. Recently, practical ML implementations of site-heterogeneous models have become available in IQ-Tree [103,104]. Frequently, subsidiary analyses are conducted to test whether initial results are robust to perturbations of the data, especially excluding data most likely to foster incorrect phylogenetic inference (e.g., the fastest-evolving species, sites, or genes).
Although pioneering phylogenomic studies were instrumental in showing what could be done, they contributed only marginally to the original supergroup model, mostly because the sampling of protist taxa was extremely limited (e.g., missing entire supergroups, especially Rhizaria) [20,26,27]. This situation gradually improved, however, and by the late 2000s some genome/transcriptome data were available for most well-known major groups [28–33]. Since then, the widespread use of next-generation sequencing, especially multiplexed transcriptomics, has greatly accelerated improvements in taxon sampling within the most familiar protist taxa [34–44]. As a result, panekaryote phylogenomic analyses of datasets of 120–350+ nucleus-encoded genes have become the dominant tool for inferring the eToL at the level of major lineages. Overwhelmingly, recent depictions of the eToL at its broadest scale are summaries of such phylogenomic analyses. Thus, unlike for the original supergroup trees, there is now little to no integration of other information (e.g., cell-biological evidence). The most important exceptions concern: (i) the placement of the root of the eukaryote tree and; (ii) inclusions of lineages known only as environmental rRNA sequences. The root is not directly examined by most phylogenomic analyses since they do not include outgroups to eukaryotes; the root must therefore be inferred using quite different data (Box 3).

New and Rediscovered Major Taxa
When the supergroup model emerged, it seemed possible that most of the major lineages had already been discovered, based on several lines of evidence. (i) Many of the former ‘mystery eukaryotes’ that had been examined using molecular tools (i.e., at least an 18S rDNA sequence was known) had been assigned to existing major groups; for example, many small flagellates and amoebae fell within Rhizaria [14]. (ii) There was only a small list of unsequenced protists that were good candidates to represent ‘new’ major lineages because they were known to have unusual cell structure [9]. (iii) Careful analyses of environmental molecular data had shown that almost all available eukaryote rRNA sequences could be assigned to known major groups [45,46]. Since 2004, however, there have been a remarkable number of new protists discovered that could be crucial for understanding the eToL at the deepest levels and a number of re-isolations of known but unsequenced taxa that have proved to be similarly important (Table 1). Almost all of these are free-living heterotrophic flagellates or amoebae. Strikingly, the great majority were isolated, and made available for molecular study, using ‘old-fashioned’ culturing approaches rather than the higher-throughput environmental molecular methods that have otherwise transformed protist diversity research (e.g. [47–49]). Most of these new protist lineages are rare or even undetected in molecular environmental data from well-studied systems such as the marine pelagic and/or include only a tiny number of known species (often just one or two). Nonetheless, the addition of these taxa to phylogenomic analyses, mostly over the past 5 years, has transformed the catalog of branches that comprise the deep-level structure of the eToL [50–63].

The Current Tree
Integrating the results of phylogenomic analyses and the main lineages added over the past 15 years, the current consensus tree has been shuffled to the extent that most of the original supergroups have either been subsumed into new taxa or disappeared altogether (Figure 1). Changes have come from three main processes: the splitting of old supergroups, often followed by reorganization in different parts of the tree; the amalgamation of isolated taxa into new larger clades; and the wholesale addition of new supergroup-level taxa. The result is a model of the eToL that has retained an overall shape similar to that of earlier broad schemes, but the details of which are vastly different. Below we briefly introduce a current listing of eukaryotic supergroups, noting the strength of evidence supporting them.

TSAR
The acronym TSAR stands for the group’s constituent members: telonemids, stramenopiles, alveolates, and Rhizaria. The latter three groups form a clade, ‘SAR’ or ‘Sar’, that emerged relatively early in the phylogenomic era [29,30,33] and has been routinely considered a ‘supergroup’ (partly replacing chromalveolates; Box 1). Sar has been estimated to comprise up to half of all eukaryote species...
diversity [4]. It includes several major groups of microbial algae (e.g., diatoms, dinoflagellates), large seaweeds (e.g., kelps), ecologically important free-living protozoa (e.g., ciliates, foraminiferans, radiolarians), and many well-studied protozoan parasites (e.g., apicomplexans, oomycetes) [64]. The sister group to Sar had been unclear, but there is now good evidence that this is the enigmatic free-living flagellate taxon Telonemia, which has just two described species [65]. The TSAR clade was robustly supported in recent phylogenomic analyses with improved sequence quality and quantity [62] and was also recovered earlier with some smaller datasets [52,53,61].

Table 1. Candidate New Major Lineages of Eukaryotes Identified since 2004 Using Molecular Phylogenetics

<table>
<thead>
<tr>
<th>Group</th>
<th>Year identified</th>
<th>Original description</th>
<th>General category</th>
<th>Molecular data source</th>
<th>Phylogenomic confirmation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Katoblepharids</td>
<td>2005 [88]</td>
<td>1939</td>
<td>Heterotrophic flagellates</td>
<td>Cultivation</td>
<td>2012 [53]</td>
</tr>
<tr>
<td>Telonemia</td>
<td>2006 [65]</td>
<td>1913</td>
<td>Heterotrophic flagellates</td>
<td>Cultivation</td>
<td>2009 [51,62]</td>
</tr>
<tr>
<td>Rigifilids</td>
<td>2008 [90]</td>
<td>2001</td>
<td>Heterotrophic amoebae</td>
<td>Cultivation</td>
<td>2018 [56]</td>
</tr>
<tr>
<td>Palptomonas</td>
<td>2010 [91]</td>
<td>2010</td>
<td>Heterotrophic flagellates</td>
<td>Cultivation</td>
<td>2014 [58]</td>
</tr>
<tr>
<td>Microheliella</td>
<td>2012 [95]</td>
<td>2012</td>
<td>Heterotrophic amoebae</td>
<td>Cultivation</td>
<td>2015 [57]</td>
</tr>
<tr>
<td>Ancoracysta</td>
<td>2017 [60]</td>
<td>2009</td>
<td>Heterotrophic flagellates</td>
<td>Cultivation</td>
<td>2017 [60]</td>
</tr>
<tr>
<td>Hermissitophora</td>
<td>2018 [61]</td>
<td>1893</td>
<td>Heterotrophic flagellates</td>
<td>Single-cell isolation</td>
<td>2018 [61]</td>
</tr>
</tbody>
</table>

*Defined as taxa that do not fall inside any robust clade within eukaryotes that was widely recognized in 2004.

*Report of cultivation and first molecular data from [97], but misidentified as an archamoeba (Amoebozoa); arguably, first identification as a likely major lineage, albeit nominally within Amoebozoa, by [98].

*Here and elsewhere, ‘cultivation’ indicates that strains have been grown indefinitely under laboratory conditions with no other eukaryotes, except prey for organisms that consume other eukaryotic cells.

*First phylogenomic investigation placed breviates incorrectly within Amoebozoa [50]. Current placement in Obazoa robustly established later [55].

*Confirmed as distinct in [51], but robust inference as sister of Sar reported much later [62].

*Named ‘picobiliphytes’ and identified as algae when first reported [89]. Later studies, including transient cultivation, show that they are heterotrophic flagellates [99,100].

*Subsequently, genome amplification performed on isolated single cells [79], these data used for seven-gene phylogenies [79] and later in phylogenomic analyses [53].

*Previously studied as ‘Micronucleariida’ [90]; current name ‘Rigifilida’ introduced in [101].

*Confirmed as distinct in [69]; robust inference of current position in CRuMs established later [56].

*Recognized as distinct, and sister to haptophytes, on the basis of plastid rDNA data only [94]. Examinations of nuclear data awaited.

*Falls outside current supergroups in phylogenomic analyses, but position is highly unstable [57]. Reanalysis awaited.

*First studied Ancoracysta was misidentified as a Colponema (Alveolata) and recognized after the fact [70]. Name introduced in [60].

*No published phylogenomic analysis, although a possible affinity with metamonads based on unpublished analyses is noted in the description [96].

*Initial small subunit (SSU) rDNA and transcriptomic data generated using single-cell methods; cultivated subsequently [61].

*Heterotrophic, but inferred to possess a nonphotosynthetic plastid based on gene sequence information [63].
Haptista

Haptista comprises the haptophyte algae (previously assigned to chromalveolates; Box 1) and centrohelids. Haptophytes, especially the calcifying coccolithophorids (e.g., *Emiliania huxleyi*), play crucial roles in marine ecosystems and global biogeochemical cycles. Centrohelids, by contrast, are free-living protozoa with ray-like pseudopodia supported by microtubules (axopodia), which radiate from a spherical cell body. Haptista is generally well supported in recent phylogenomic studies [54,57].

Cryptista

Cryptista contains the cryptomonads (also former chromalveolates; Box 1), a lineage that has been central to the study of plastid origin and spread across eukaryotes (e.g., *Guillardia theta*). Cryptista also includes the katablepharids and the more recently discovered *Palpitomonas*, both enigmatic heterotrophic flagellates (Table 1). Phylogenomic studies robustly support the monophyly of Cryptista [37,53,58].

Archaeplastida

The three taxa that comprise Archaeplastida are the Chloroplastida (green algae + land plants), Rhodophyta (red algae), and Glaucophyta. All three lineages have primary plastids, which are photosynthetic organelles that originated directly from cyanobacteria. Recently, a new group – *Rhodelphis* – was discovered and shown to branch as sister to red algae in phylogenomic analyses [63]. *Rhodelphis* cells are heterotrophic flagellates, but gene sequence data suggest that they have a
Box 3. The Root of the eToL

Although the position of the root is fundamental to our understanding of the eukaryotic tree, it is usually not addressed by phylogenomic analyses aimed at resolving deep branches within the eToL. In the early 2000s, the presence or absence of several discrete molecular and cell-biological properties in various supergroups was used to argue that the root of the eukaryotic tree fell between two major clades: the ‘unikonts’ (including Opisthokonta and Amoebozoa; i.e., Amorphea) and the ‘bikonts’ (almost all other eukaryotes) [105,106]. Unfortunately, as molecular data from diverse protists became available, the distribution across taxa of these discrete molecular markers no longer cleanly supported a unikont/bikont root [107–109]. Subsequent studies using several approaches other than pure molecular phylogenetics have inferred a variety of potential eukaryote root placements including: (i) between Archaeplastida and all other eukaryotes [110]; (ii) between Opisthokonta and other eukaryotes [66]; (iii) jakobids versus other eukaryotes [30]; and (iv) the Euglenozoa versus other eukaryotes [111].

More recently, the root position has been addressed using molecular phylogenies of concatenated proteins of mitochondrial or bacterial origin in which eukaryotes appear particularly closely related to outgroup prokaryotic sequences [112–114]. Derelle and Lang [112] analyzed 42 genes of mitochondrial origin and found that their analyses supported a ‘unikont’/’bikont’ root. Then, He et al. analyzed a distinct, but overlapping, set of 37 genes, including some transferred to the eukaryotic stem lineage from bacteria prior to the Last Eukaryote Common Ancestor (LECA) [114]. Their analyses placed the eukaryote root on the branch between Discoba and other eukaryotes. Derelle and colleagues subsequently contested this result, recovering a root between two large groupings: ‘Opimoda’, comprising Amorphea, collocidiontids, and malawimonads, and ‘Diphoda’, including Discoba, Archaeplastida, cryptomonads, and Sar [113]. They argued that Excavata cannot be a natural group because both ‘sides’ of this root include excavates (malawimonads and Discoba, respectively). If correct, this root implies that cell-structure features proposed as synapomorphies for Excavata could be ancestral properties of the LECA. Regardless of which, if any, of these results are correct, many of the novel protist taxa recently placed in the eToL (Figure 1 and Table 1) were not represented in these analyses. Therefore, the precise position of the root of the eToL remains uncertain.

Amorphea

This taxon groups opisthokonts (animals, fungi, and their respective unicellular relatives) with the amoeboid protists of Amoebozoa (e.g., Amoeba and most ‘slime molds’ among many). Amorphea now also includes two small lineages of heterotrophic flagellates, the breviates and the apusomonads, that cluster with the opisthokonts to form the Obazoa [34,55]. Amorphea is robustly supported in most phylogenomic analyses, with the caveat that the position of the root remains uncertain (Box 3), and a placement within Amorphea has been inferred in some cases [66], which would make Amorphea paraphyletic.

CruMs

As with TSAR, CRuMs represents a novel proposed supergroup named as an acronym of its constituent members: collocidiontids (syn. diphyleids) + Rigifilida + Mantamonas. These three free-living protozoan taxa have very different basic morphologies (swimming flagellates, filose amoeboid cells, and tiny gliding cells, respectively) and were previously ‘orphan taxa’ (see below), but robustly coalesced in recent phylogenomic analyses [56,61].

Discoba

Discoba includes Euglenozoa and Heterolobosea (collectively ‘Discicristata’), plus the heterotrophic flagellate groups Jakobida and Tsukubamonas (Table 1). Euglenozoa includes the euglenophyte
algae, trypanosomatid parasites, and numerous free-living or parasitic heterotrophic flagellates, which are abundant in many ecosystems. Heterolobosea are heterotrophic amoebae and flagellates. Discoba was suspected on the basis of selected single- and multigene phylogenies and strongly confirmed by phylogenomic analyses [31,59].

**Metamonada**

Metamonada entirely comprises anaerobic protists, including various free-living protozoa, intestinal symbionts (especially of wood-eating insects), and many parasites (e.g., Giardia, Trichomonas). The monophyly of Metamonada is well supported by contemporary phylogenomic analyses [38,67]. However, placing metamonads relative to other taxa has proved very challenging, because most species exhibit very high rates of sequence evolution. Phylogenomic analyses often infer a Metamonada plus Discoba clade (see above) [34,68,69] largely corresponding to the original ‘Excavata’ supergroup (Box 1); however, this topology could represent an analysis artefact. Some phylogenomic analyses, usually those that include shorter-branching metamonads, recover instead a specific relationship with the ‘orphan’ excavate group malawimonads (see below) [55,59,68,70].

**Hemimastigophora**

The ‘hemimastigotes’ are free-living protozoa with two rows of flagella. They had been known since the 19th century and given a high taxonomic rank based on electron microscopy observations [71] but were never cultivated, and genetic data were lacking. Recent phylogenomic analyses, based on transcriptomes from hand-picked cells of two genera, showed hemimastigotes as one of the deepest branches within eukaryotes [61]. They could not be placed as sister to any one of the ‘established’ supergroups (or any ‘orphan’); consequently, it was proposed to consider them a new supergroup.

**Orphan Taxa**

In addition to the groups listed above, there are several seemingly species-poor taxa for which phylogenomic analyses have thus far failed to provide a convincing phylogenetic placement. These so-called ‘orphan taxa’ include Ancoracysta, Picozoa, malawimonads, and ancyromonads (= planomonads), all of which are free-living protozoa. Some or all of these may branch with an established group; for example, Ancoracysta may be sister to Haptista [60,62] and malawimonads may be sisters to Metamonada (see above; [68]). It is possible, however, that some represent even deeper-diverging lineages, following the recent example of Hemimastigophora.

Given this new framework for eukaryote evolution (Figure 1), an obvious question is: can these supergroups be reliably grouped further? Most recent phylogenomic analyses show Cryptista branching with (or within) Archaeplastida and many show Haptista as a close relative of Sar, and now TSAR [37,53,54,61,62]. ‘Diaphoretickes’ is an even larger assemblage that is proposed to unite these four supergroups to the exclusion of Amorphea, Discoba, and Metamonada [13,56,61,72], while CRuMs is inferred to be sister to Amorphea [56]. It is too early to tell, but even if reliable, these inferences depend on assumptions about the position of the root of eukaryotes (Box 3), which becomes ever-more problematic as larger groups are inferred from unrooted phylogenetic trees.

**The Nature of the Supergroup Model**

In addition to the list of supergroups changing greatly over the past 15 years, the typical nature of those groupings has also changed, with important consequences for how we conceptualize the supergroup model to describe the tree. As mentioned earlier, most of the original supergroups were distinguished by some conspicuous biological feature (Box 1). By contrast, the new supergroups mostly reflect clades in phylogenetic trees that lack candidate shared derived characteristics (Figure 1). For instance, under the original framework all secondary algae with red alga-derived plastids were assigned to Chromalveolata, following the assumption that those plastids were acquired in a common ancestor [73]. Today, however, the chromalveolate hypothesis is not widely accepted (although see [70] for a different opinion), largely because most phylogenetic analyses do not show close relationships among the host lineages [74]. None of the new groupings resulting from
the disintegration of Chromalveolata (TSAR, Cryptista, Haptista) is likely to be ancestrally defined by red secondary plastids [55]. A similar observation can be made for Opisthokonta, which remains distinguished by ancestrally having a single, posterior flagellum in motile cells but is now usually considered a subtaxon within Obozoa and Amorphea, neither of which has unifying morphological characteristics [13,55]. The supergroups that are still distinguished by a biological property are among the most unstable, at least with their current compositions. Archaeplastida is defined by the presence of primary plastids but remains poorly supported by phylogenetic analyses (e.g. [37,54,62]). Similarly, Metamonada consists entirely of anaerobes but is likely to be subsumed into a more inclusive taxon once the relationships among ‘excavates’ are better understood.

Decisions about which major groupings are considered supergroups have always been arbitrary, but the increasing absence of distinguishing biological features makes this more apparent. Paradoxically, the improved resolution of the tree makes the problem worse, not better. To illustrate this issue, take the newly identified supergroup CRuMs [56], which was inferred to branch together with Amorphea, itself containing two taxa often recognized as supergroups, Amoebozoa and Obozoa. The opinion that this collection of taxa represents two supergroups (or three), rather than one, reflects the lack of distinguishing characters for the CRuMs–Amorphea grouping. This leaves the decision driven by subjective judgments concerning: (i) which phylogenetic results are sufficiently robust to be accepted without further confirmation; and (ii) the uncertainty about the location of the ‘root’ of the eToL (Box 3). Moreover, there is a blurry line between orphan lineages, which often have just a few known species, and the least speciose supergroups. If a diversity-poor orphan is shown to be evolutionarily unrelated to all supergroups, does that make it a new supergroup? To be most useful, the notion of ‘supergroup’ should not be distinguished from ‘orphan’ by the level of diversity it contains but instead should reflect the degree of confidence that a lineage is not encompassed phylogenetically by an existing clade.

We expect that many researchers and educators will continue to find it useful to divide eukaryote diversity into a small number of major clades, and this ultimately is what a catalog of supergroups aims to provide. Future comparative genomics research may identify robust apomorphies for deep clades within eukaryotes, which in turn could help to more naturally delineate supergroups. Until then, however, it seems that the bulk of major subdivisions of eukaryotes will continue to be only clades derived from molecular phylogenetic trees. Accordingly, we should expect the list of supergroups to be increasingly volatile as the understanding of eukaryote diversity and resolution of the tree improve further (see below), and more author-dependent, since there will be no conspicuous criteria for deciding which clades are to be distinguished as supergroups.

Where Are We Going?
The recent discoveries of several very deep branches in the eToL have enabled a profound re-evaluation of major evolutionary transitions that occurred hundreds of millions of years ago (e.g. [58,60,61]). Remarkably, these organisms were mostly identified using low-throughput, classical culturing approaches, and the rate of such discoveries shows no sign of tailing off (Table 1). In parallel, however, we are seeing a rapid maturation and greatly increased accessibility of single-cell transcriptomic and genomic methods, which do not rely on culturing [40,41,61,75–77]. Large numbers of cells can be isolated en masse from the environment then screened using molecular techniques to identify important organisms for further study [78–83]; alternatively, target cells can be identified by microscopy and selected individually [40,41,61,75]. Both single-cell genomics and transcriptomics use amplification techniques and typically generate relatively incomplete assemblies and biased representation. Nonetheless, the transcriptomic approach, at least for larger cells, can yield very good coverage in phylogenomic datasets, which are dominated by high-expression housekeeping genes [41,61].

Further development of systematic, higher-throughput methods to explore the microbial eukaryote fraction of ecosystems is also needed. Metagenomics (or metatranscriptomics) has revolutionized research on prokaryotes, providing thousands of reconstructed genomes for organisms that may
never have been observed under a microscope. Eukaryotic metagenomics is still in its infancy, but the signs are there that it might be a workable approach for placing novel genetic diversity in a phylogenomic framework [84,85]. Another method recently applied to obtain genomic information from important taxa combines metabarcoding and fluorescence in situ hybridization (FISH) to go from sequences back to the cells [86]. No matter which technique proves to be the most useful, the release from the burden of culturing means that the taxonomic breadth, and importantly the taxon density, of phylogenomic datasets may improve rapidly in the near future. Thus, new groups that are especially challenging to culture may be identified and added for the first time, in turn greatly accelerating the achievement of robust taxon sampling for all groups on the tree. The availability of these data should ultimately improve the overall reliability of phylogenetic estimation, although with the caveat that using culture-free approaches generally means that some important aspects of the biology are missed (e.g., details of life cycles and morphology).

With these anticipated improvements in taxon sampling for eukaryotes, it is more important than ever to develop rigorous phylogenomic pipelines. This involves best practice when assembling the datasets as well as models of sequence evolution complex enough to adequately describe the processes at play, with software implementations that allow these models to be used on large datasets (Box 2). So far, broad-scale phylogenomics of eukaryotes has almost exclusively used the concatenation approach, but exploring, in depth, the influence of individual genes can help to pinpoint more specifically where and how the phylogenetic signal is distributed [55,87]. It will also be informative to assess the origins of the different signals between different datasets so that the influence of taxon and gene sampling can be disentangled. Better understanding of eukaryote-wide phylogenomic datasets, combined with improvements in state-of-the-art phylogenetic methods, will enable the recovery of even more ancient and difficult-to-discern phylogenetic signals.

Concluding Remarks and Future Perspectives

The eToL has been considerably remodeled in the past 15 years following the development of phylogenomics and the addition of evolutionarily key protist taxa. The support for the major groups has shifted from being a synthesis of various molecular phylogenetic evidence and biological characters to being based almost entirely on multigene molecular phylogenies. An indirect but important consequence of this shift is that it is increasingly difficult to describe the tree in simple terms, although the resolution of the tree itself has improved greatly and continues to do so. The eToL will always be a ‘work in progress’ and with the incremental changes over time our understanding of evolutionary relationships advances. With the maturation of the phylogenomic approach, we are better equipped than ever before to improve the resolution of the eToL and to facilitate its interpretation in light of the unprecedented amount of data generated for a growing diversity of protists (see Outstanding Questions).

Acknowledgments

F.B.’s research is supported by a Fellowship grant from SciLifeLab, a research project grant from the Swedish Research Council (VR-2017-04563), a Future Research Leader grant from Formas (2017-01197), and a postdoctoral fellowship from Carl Tryggers Stiftelse. A.J.R. and A.G.B.S. are supported by Discovery Grant 2016-06792 and 2019-298366, respectively, from the Natural Sciences and Engineering Research Council of Canada (NSERC). M.W.B.’s research is supported by the US National Science Foundation (NSF) Division of Environmental Biology (DEB) grant 1456054.

References


Outstanding Questions

How many more extant ‘kingdom-level’ eukaryotic lineages exist and can we find them? Most recent discoveries of major groups used culture-based approaches, whilst much higher-throughput environmental sequencing has mostly increased the diversity within known supergroups. Interestingly, the major lineages discovered via cultivation are often not well represented in environmental surveys; this suggests that cultivation and culture-independent methods will preferentially access different subsets of the diversity remaining to be characterized.

Can we refine the relationships among the major lineages to obtain a fully resolved eukaryotic Tree of Life (eToL)? Will phylogenomics using more deep-branching taxa and better evolutionary models (e.g., site-heterogeneous models) be enough to stabilize all major nodes in the eToL?

Can we find apomorphic characters that support the phylogenetically derived supergroups? Traditionally, these characters were cell-biological features. With more genomes sequenced across the full breadth of eukaryote diversity, supporting characters may well be found at different organizational levels; for example, as genomic innovations like gene gains and losses.

What is the relative importance of gene-sampling and taxon-sampling to recovering the deep branches in the tree? What is the minimal number of genes required? Which taxa are the most fundamental to stabilize, or disturb, eukaryote-wide phylogenies?

What is the timescale of eukaryote evolution and under what ecological conditions did the major transitions happen? The most recent molecular estimates of divergence times among all eukaryotes were based on previous versions of the eToL. The new tree, and the much denser taxon sampling now available, provides the opportunity to address more precisely the timing of major evolutionary events with respect to geological history.


68. Heas, A.A. et al. (2018) Combined morphological and phylogenomic re-examination of alveolinids, a critical taxon for inferring the evolutionary history of eukaryotes. R. Soc. Open Sci. 5, 171707


82. 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


99. Seenivasan, R. et al. (2013) Picomonas judraskeda gen. et sp. nov.: the first identified member of the Picozoza phylum nov., a widespread group of picoeukaryotes, formerly known as "picobiliphytes". PLoS One 8, e59565
110. Rogozin, I.B. et al. (2009) Analysis of rare genomic changes does not support the unikont–bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. Genome Biol. Evol. 1, 99–113