Microbial Diversity and Open Questions about the Deep Tree of Life

Laura Eme [1,*] and Daniel Tamarit [2,*]

[1]Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif sur-Yvette, France
[2]Theoretical Biology and Bioinformatics, Utrecht University, Utrecht 3584CH, The Netherlands

*Corresponding authors: E-mails: laura.eme@universite-paris-saclay.fr; d.tamaritchulia@uu.nl.

Accepted: March 11, 2024

Abstract

In this perspective, we explore the transformative impact and inherent limitations of metagenomics and single-cell genomics on our understanding of microbial diversity and their integration into the Tree of Life. We delve into the key challenges associated with incorporating new microbial lineages into the Tree of Life through advanced phylogenomic approaches. Additionally, we shed light on enduring debates surrounding various aspects of the microbial Tree of Life, focusing on recent advances in some of its deepest nodes, such as the roots of bacteria, archaea, and eukaryotes. We also bring forth current limitations in genome recovery and phylogenomic methodology, as well as new avenues of research to uncover additional key microbial lineages and resolve the shape of the Tree of Life.

Key words: Tree of Life, microbial diversity, microbial genomics, phylogenomics, metagenomics, single-cell omics, artificial intelligence.

Significance

The emergence of metagenomics has profoundly transformed our understanding of microbial diversity. In the last decade or so, this innovative approach has led to the discovery and substantial expansion of numerous microbial groups through the reconstruction of their DNA sequences. This perspective delves into how recent developments in phylogenomic methods are facilitating the identification of these new lineages within the Tree of Life. Additionally, we explore how incorporating this increased diversity clarifies some of the oldest branches of this tree. We conclude by highlighting key research directions that are poised to greatly advance our understanding of the evolutionary history of life.

This perspective is part of a series of articles celebrating 40 years since our sister journal, *Molecular Biology and Evolution*, was founded. The perspective is accompanied by virtual issues on this topic published by *Genome Biology and Evolution* and *Molecular Biology and Evolution*, which can be found at our 40th anniversary website.

Expanding Microbial Diversity through Metagenomics and Single-Cell Approaches

The advent of environmental sequencing and metagenomics has revolutionized our understanding of microbial diversity on a monumental scale. These powerful techniques have enabled scientists to delve deep into the invisible and complex world of microorganisms that inhabit our planet (Venter et al. 2004; Hugenholtz and Tyson 2008). By directly analyzing genetic material from environmental samples such as soil, water, or the human gut, researchers regularly uncover an astonishing array of previously unknown microorganisms. Metagenomics, in particular, allows for the simultaneous study of myriad species in their natural habitats, providing insights into their roles in ecosystems and their contributions to global nutrient cycles

(López-García and Moreira 2008; Grossart et al. 2020). Thanks to these technologies, our appreciation of the vast and intricate microbial communities that underpin life on Earth has expanded exponentially, with profound implications for the field of evolutionary biology.

While metagenomics is a powerful tool for exploring microbial diversity and functionality, it is not without its limitations. One significant constraint lies in the fact that metagenomics has typically relied on short DNA sequences, which can make it challenging to assemble complete genomes and accurately determine the gene repertoire and metabolic potential of individual species within a community. The recent emergence of long-read metagenomics holds great promise, but it is still in its infancy, notably due to the large quantity of high-molecular-weight DNA necessary, the higher error rate compared to short-read sequencing, and its high cost caused by the high coverage needed to offset the error rate. Another limitation is the risk of contamination and biases in metagenomic data. Environmental samples can easily become contaminated with DNA from other sources, leading to false interpretations of community composition. Furthermore, the choice of DNA extraction protocol, sequencing technology, and data analysis methods can introduce biases, impacting the representation of certain microbial groups.

In particular, the application of metagenomics to study unicellular eukaryotes (protists) faces several inherent drawbacks. This is primarily due to the often low abundance of protists in environmental samples, combined with the sheer size and complexity of eukaryotic genomes compared to prokaryotic ones. The genetic material of eukaryotes generally contains numerous noncoding regions, repetitive elements, and introns, making it more difficult to sequence in-depth and completely, and to assemble and bin the sequencing data accurately. Moreover, this low abundance can lead to incomplete or biased representation of the community. Consequently, metagenomics may not fully capture the diversity and genetic potential of microbial eukaryotes, necessitating targeted approaches, such as single-cell genomics.

Single-cell genomics has emerged as a powerful approach for studying protists, offering several advantages and overcoming significant challenges. One of the key advantages is the ability to isolate and analyze the genetic material from individual protist cells, allowing to place them in the Tree of Life, and for a more precise understanding of the genomic diversity within complex microbial communities while limiting cross-contamination issues. This approach is particularly valuable when dealing with protists that are difficult to culture or that can simply not be established in pure culture as they often feed on other microorganisms. Additionally, single-cell genomics (or transcriptomics) can provide insights into the functional potential of these organisms by identifying key genes and metabolic pathways. Recent examples of protist sequence data obtained through single-cell (or "few-cell") approaches include the discovery of novel high-ranking lineages (Lax et al. 2018; Wideman et al. 2020; Schön et al. 2021). Moreover, single-cell genomics has enabled the study of symbiotic and parasitic protists, shedding light on their genetic adaptations and evolutionary relationships (Dia and Cheeseman 2021; Boscaro et al. 2023).

The ToL: A Few Knowns, Many Unknowns

The expansion of environmental sequencing data encompassing a more extensive range of Earth's true microbial biodiversity, which constitutes the majority of life, has led to a significant enrichment and refinement of the Tree of Life (ToL). For instance, in just the last 4 years, the number of identified bacterial families and orders in the Genome Taxonomy Database (Parks et al. 2022) has doubled and entirely new high-ranked taxa have been discovered very recently. A decade ago, no genomic data existed for significant groups like the Asgard archaea (Spang et al. 2015) and other highly diverse groups, such as Patescibacteria (a.k.a. Candidate Phyla Radiation, CPR bacteria), DPANN archaea, or the Planctomycetes–Verrucomicrobia–Chlamydiae bacterial clade, were known only through a handful of classical representatives (Rinke et al. 2013). This surge in sequenced diversity enriches the breadth of phylogenetic signals available, greatly enhancing our ability to reconstruct deep relationships in the ToL and providing a more comprehensive understanding of life's evolutionary history.

However, placing the myriad of newly discovered microbial lineages in the ToL remains a complex task with several limitations (Kapli et al. 2021). A common approach to infer species trees relies on combining the phylogenetic signal carried by multiple proteins, either through a concatenation or a supertree approach, which are thought to amplify the often weak information contained in each individual marker.

One of the most prominent challenges linked to these approaches is the occurrence of nonvertical signals in individual gene phylogenies, blurring reconstruction of the species tree. This signal can be biological and reflect the true history of the gene, or artifactual due to errors in the data or limitations in the phylogenetic methods employed (Philippe et al. 2017; Steenwyk et al. 2023). One common biological source of nonvertical signal is horizontal gene transfer, where genes are exchanged between unrelated species. This phenomenon is prevalent in bacteria and archaea (Zhaxybayeva and Doolittle 2011) and commonly reported in microbial eukaryotes (Andersson 2009; Sibbald et al. 2020), and complicates the determination of evolutionary relationships. Consequently, a critical aspect of phylogenomic approaches is the accurate identification and curation of orthologs—genes in different species that

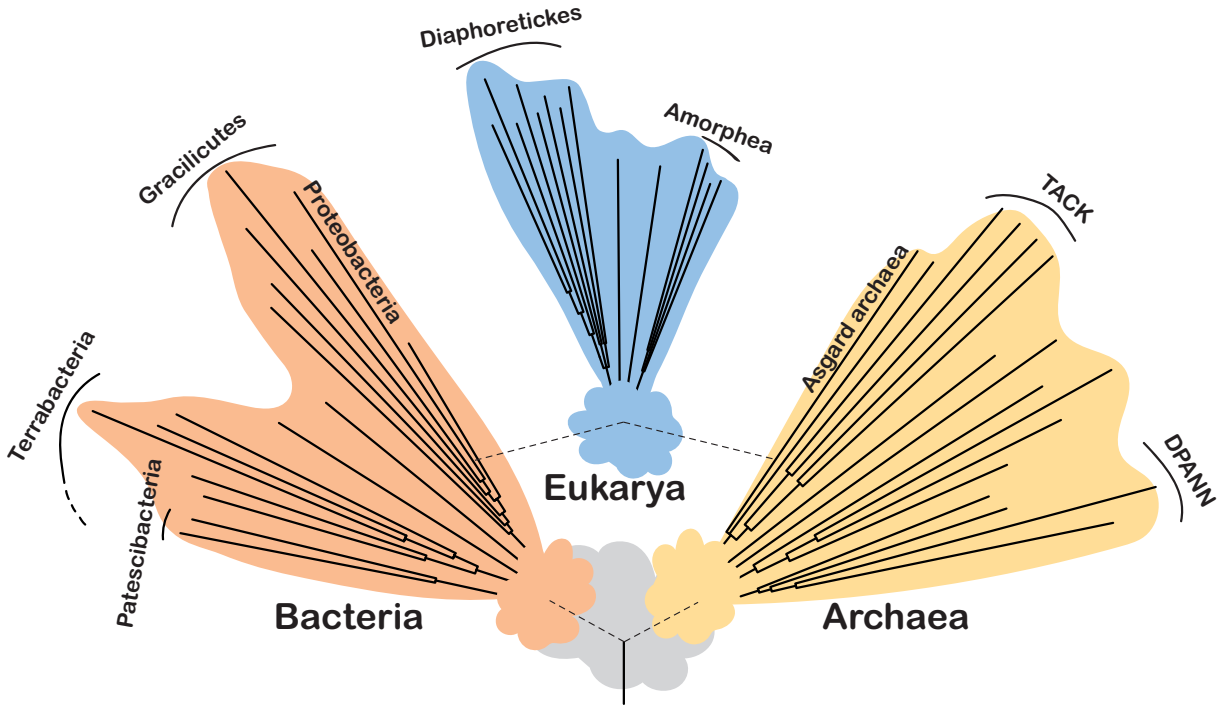

Fig. 1.—Schematic depiction of the uncertainty related to some key branches of the ToL discussed in this work.

originated from a common ancestral gene through speciation, and whose history mirrors the one of the organisms that carry them. The development of sophisticated computational tools and databases has been instrumental in facilitating the identification and curation of orthologs, thereby enhancing the reliability and precision of phylogenomic studies (Tice et al. 2021; Richter et al. 2022; Comte et al. 2023; Hernández-Plaza et al. 2023).

In addition, the specification of the evolutionary model in phylogenetic analyses is a critical step that can significantly influence the results and conclusions drawn about the relationships among species. Choosing an appropriate model that accurately reflects the evolutionary processes (e.g. amino acid substitution patterns and rate variations across sites and lineages) is crucial. A model that does not properly capture the true evolutionary process will risk yielding incorrect phylogenies. A common artifact due to model misspecification is known as long-branch attraction (LBA), which corresponds to rapidly evolving lineages (long branches, such as those corresponding to parasitic lineages) being erroneously inferred to be closely related, regardless of their true relationship (Felsenstein 1978; Susko and Roger 2021). LBA is particularly problematic in cases where the model of evolution used does not properly account for rate differences across lineages. Moreover, incorrect model specification can affect branch length estimations as well as the inference of ancestral states, potentially leading to erroneous conclusions about the timing of evolutionary events and the nature of ancestral species (Susko et al. 2021; Del Amparo and Arenas 2022). Therefore, careful model selection, guided by statistical methods and biological understanding of the lineages under scrutiny, is essential to ensure the reliability and accuracy of phylogenetic reconstructions.

Similarly, the most commonly used models of evolution do not account for changes in amino acid preferences that can occur across the tree (i.e. between lineages), generally due to an evolutionary transition to a new lifestyle or environment. The adaptation of microbial organisms to extreme environments, such as high-temperature or high-salt conditions, often results in shifts in amino acid preference in their proteins (De Farias and Bonato 2002; Fukuchi et al. 2003; Reed et al. 2013; Eme et al. 2023). For instance, in hyperthermophilic organisms, there is a preference for amino acids that contribute to increased protein stability at high temperatures, such as amino acids that can form more ionic bonds. Similarly, halophilic organisms often have proteins with a higher proportion of acidic amino acids, allowing them to remain functional in high-salt environments. These adaptations pose challenges for phylogenetic reconstructions as these atypical patterns of molecular evolution that are found only in specific parts of the tree, when not properly accounted for, can result in phylogenetic artifacts. The accelerated substitution rates or convergent amino acid substitutions in these organisms might be misinterpreted as close evolutionary relationships between unrelated lineages. Therefore, when including such extremophilic lineages in phylogenetic analyses, it is essential

to employ models that can accommodate these distinctive evolutionary patterns (Williams et al. 2021; Muñoz-Gómez et al. 2022; Baker et al. 2024).

Altogether, this highlights the importance of developing more realistic evolutionary models; this is currently a frontier in evolutionary (micro)biology (Susko and Roger 2007, 2020; Susko et al. 2018).

Persistent Controversies in the ToL

In part due to all the aforementioned limitations, many areas of the ToL remain challenging to resolve (Fig. 1). The prevailing view in evolutionary biology is that Bacteria and Archaea, each as a monophyletic group, share a history that traces back to the last universal common ancestor (LUCA). LUCA stands as a pivotal entity in our understanding of life's history, representing the most ancient organism for which we can methodically infer the genome content using comparative genomics and phylogenetics. It serves as a critical juncture, linking the known biosphere with the enigmatic origins of terrestrial life. However, our understanding of LUCA's characteristics is deeply intertwined with its hypothesized position in the ToL. Consequently, accurately determining the phylogenetic relationship between Bacteria and Archaea is crucial, as it directly influences our inferences about LUCA's nature and the early evolutionary pathways that shaped life on Earth. One inherent difficulty lies in the fact that the ToL's root cannot be determined using outgroup sequences. Decades ago, researchers aimed to palliate this issue by analyzing protein families that duplicated before LUCA, allowing paralogs to act as each other's outgroup (Gogarten et al. 1989; Iwabe et al. 1989; Brown and Doolittle 1995). This methodology predominantly suggested a root between the bacterial and archaeal domains. Alternative approaches, such as analyses of ribosomal protein composition (Fournier and Gogarten 2010) and gene presence–absence (Dagan et al. 2010), also supported a root placement between Bacteria and Archaea. Alternative hypotheses, proposing roots within the bacterial domain or between eukaryotes and prokaryotes (Brinkmann and Philippe 1999; Lopez et al. 1999), have been suggested but lack robust corroboration, especially in light of the expanded microbial diversity. More recently, a novel phylogenetic "cross-bracing" approach was consistent with a root between Bacteria and Archaea, but not conclusive (Mahendrarajah et al. 2023). Surprisingly, this question has not been thoroughly revisited in a long time. We suspect that phylogenetic approaches such as gene–tree–species–tree reconciliation might provide further insights into this complex and long-standing question in evolutionary biology (Williams et al. 2023).

Inferences about the nature of LUCA also depend on the shape of the rest of the ToL, including the placements of the bacterial and archaeal tree roots, both of which have dramatically shifted with the astounding microbial diversity discovered in the last decade. Patescibacteria (CPR bacteria) and DPANN archaea, two large phylum-level groups of bacteria and archaea, respectively, represented by nanosized cells and small genomes, have played a critical role. Multiple studies that used Archaea as an outgroup have suggested a bacterial root between the Patescibacteria and other bacteria (Hug et al. 2016; Zhu et al. 2019). Similarly, reconstructions of the archaeal tree using a bacterial outgroup have suggested that the root might be within Euryarchaea or between DPANN and the rest of Archaea (Williams and Embley 2014). However, here again, phylogenetic reconstructions face significant challenges due to long branches leading to Patescibacteria and DPANN, and the ultralong branch separating Bacteria and Archaea, which increase the risk of LBA artifacts. Moreover, besides fast-evolving lineages, some DPANN lineages have proteomes that are compositionally adapted to extremely high-salt conditions and whose inclusion in phylogenies requires specific modeling (Baker et al. 2024). Consequently, the root positions of Archaea and Bacteria are still considered unknown, and some studies have attempted to minimize reconstruction artifacts through outgroup-free analyses. These have provided additional support for an archaeal root between the DPANN group and the rest of Archaea (Williams et al. 2017) and a bacterial root between two major groups comprising most of the known bacterial diversity, Gracilicutes and Terrabacteria (Coleman et al. 2021).

Moving closer to our time, the last eukaryotic common ancestor (LECA) marks the culmination of eukaryogenesis, a complex evolutionary process through which the eukaryotic cell emerged from at least two prokaryotic ancestors: an Asgard archaeon and an alpha-proteobacterial (mitochondrial) symbiont (Fig. 1). The large divergence time between LECA and these ancestors (Betts et al. 2018; Mahendrarajah et al. 2023) and the changes in the evolutionary mode and tempo over the branches leading to LECA represent severe impediments to accurately infer our closest prokaryotic ancestors. Here, beyond the massive expansion in available genomic data both for Archaea and Bacteria, advances have come from extensive phylogenomic studies combining complex evolutionary models, careful treatments of compositional sequence biases, and manual data curation. These have indicated a deep position of the mitochondrial branch within alpha-proteobacteria (Martijn et al. 2018; Muñoz-Gómez et al. 2022) and a placement of the nucleocytoplasmic branch within the Asgard archaeal group Heimdallarchaeia (Eme et al. 2023). In the coming years, these results will be corroborated or contested by further characterization of microbial diversity and the application of sophisticated phylogenomic methods that consider the long stem branches of eukaryotes and the complex evolutionary dynamics of these groups.

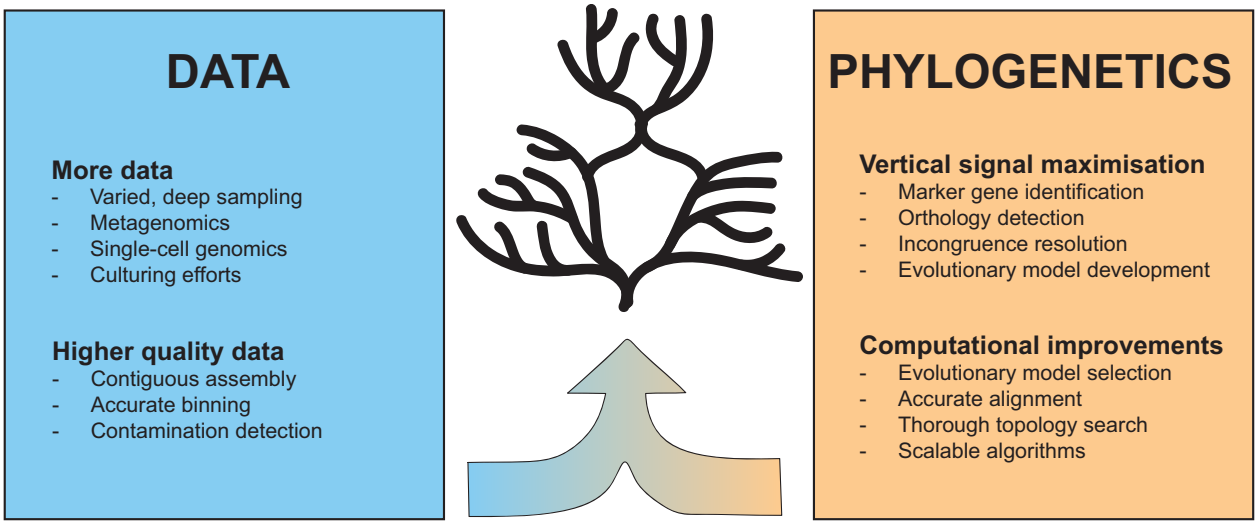

Fig. 2.—Axes of improvement for the accurate reconstruction of the ToL.

Finally, the position of the LECA itself remains an open question. While the majority of known eukaryotic species are multicellular (animals, land plants, and fungi), these account for only a fraction of the overall diversity across eukaryotes. The vast majority of eukaryotes are protists. Understanding the deepest part of the eukaryotic ToL (eToL) thus equates to resolving the relationships among major microbial lineages. Over the past decade, the deep structure of the eukaryotic tree has undergone substantial revisions, propelled by breakthroughs in phylogenomics and the integration of numerous evolutionarily pivotal protist lineages into molecular studies (Brown et al. 2018; Lax et al. 2018; Burki et al. 2020; Tikhonenkov et al. 2022). However, clarifying the earliest divergences in the eToL, including the location of its root, remains a formidable challenge. This difficulty arises partly because some lineages, known as "orphan lineages," are so sparsely represented by sequence data that their phylogenomic placement remains ambiguous. Additionally, it is believed that eukaryotes rapidly diversified from their last common ancestor over a billion years ago, giving rise to most supergroups in a brief time frame (Eme et al. 2014; Betts et al. 2018; Mahendrarajah et al. 2023). This rapid diversification has left scant phylogenetic evidence to identify relatedness between these groups. Another significant obstacle is the vast evolutionary gap between eukaryotes and their closest prokaryotic relatives. Using a prokaryotic outgroup in this context is likely to introduce LBA artifacts, as in other parts of the tree. Moreover, the origins of most eukaryotic genes cannot be definitively traced back to either Asgard archaea or alpha-proteobacteria, further complicating the selection of an appropriate outgroup for these analyses (Pittis and Gabaldón 2016). As a result, numerous conflicting hypotheses have emerged regarding the eukaryotic root position, based on molecular phylogenies of concatenated proteins (Derelle and Lang 2012; He et al. 2014; Cerón-Romero et al. 2022; Al Jewari and Baldauf 2023). Additionally, many of these studies do not include the latest high-ranking protist taxa and do not always adequately mitigate artifactual signals. This highlights the necessity of incorporating a broader spectrum of taxa into these studies and enhancing methods and models to reduce phylogenetic artifacts.

Perspectives

It is surely evident from the above that improving our understanding of microbial life, from its diversity to the structure of the ToL, will come from four intertwined fronts (Fig. 2): increasing our taxon sampling through not only more but also higher quality sequence data (Fig. 2, blue frame), and bettering our phylogenetic and computational approaches (Fig. 2, orange frame). Metagenomics has led to significant breakthroughs regarding the former, and we can still expect significant advances in the coming years. Novel microbial lineages might be found in low abundances, in close association with their physical surroundings, or be endemic to secluded environments. Identifying such microbes would benefit from sampling campaigns thoroughly targeting new or understudied environments (e.g. Baker et al. 2024). Moreover, culturing remains an invaluable, if only immensely challenging, approach to generating sequence data from microbial eukaryotes, giving access to renewable sources of biological material for in-depth observations and sequencing. However, culturing novel protists can be difficult as many require often-unknown specific environmental conditions, nutrients, symbiotic partners, or microbial prey for growth. Nevertheless, advances in culturing and single-cell isolation techniques can lead to a deeper

understanding of the biology of novel protists and open doors to innovative research in genomics and microbial ecology (Lax et al. 2018; Galindo et al. 2019; Gawryluk et al. 2019; Tikhonenkov et al. 2022). Similarly, painstakingly long cultivation efforts for prokaryotes from various environments for which we knew very little have proven to be invaluable, as shown by the recent cultures of Asgard archaea (Imachi et al. 2020; Rodrigues-Oliveira et al. 2022). These first representatives have allowed us to confirm some of the phenotypical properties of these organisms that were predicted based on their genomes alone and also provide crucial insights into their ecology and cell biology.

Equally important to the future of understanding the ToL is the improvement of phylogenetic methods. Enhancements in phylogenetic models better taking into account heterogeneity of the substitution process, composition, and evolutionary rates across sites and over time, as well as the development of new frameworks such as gene–tree–species–tree reconciliation, will allow for more accurate reconstructions of evolutionary histories. However, the use of complex evolutionary models is computationally intensive, making their usage near impossible on data sets containing several hundreds or thousands of species. Thus, another avenue of progress will be to combine the improved modeling of protein evolution with divide-and-conquer approaches, which are computational strategies designed to manage the analysis of large and complex data sets. They break down a large phylogenetic problem into smaller, more manageable parts, analyze these parts separately, and then combine the results to reconstruct a final phylogenetic tree or to update an existing tree with numerous taxa (Balaban et al. 2023).

However, alongside these technological advancements, there is a growing recognition of the importance of thorough curation of data sets. The idea that analyzing more data is the key to solving the relationships between organisms has not always proven to be true, as data quality often trumps quantity, especially when the latter includes noise and artifactual signal.

Furthermore, there is a renewed emphasis on corroborating phylogenetic hypotheses with synapomorphies—shared derived characters that support a particular evolutionary topology. For example, structural analyses of the ribosome were the first indications of the archaeal ancestry of the eukaryotic informational machinery, leading to the proposal of the eocyte hypothesis (Lake 1985). More recently, the discovery that Asgard archaeal genomes encoded more eukaryotic signature proteins than other prokaryotes brought additional support to their unique evolutionary relationship with eukaryotes (Spang et al. 2015; Eme et al. 2023). Similarly, a number of unexpected features linked to the emergence of multicellularity have been found to support the relationship between animals and their closest unicellular relatives (Ruiz-Trillo et al. 2023). This underscores the importance of a multifaceted approach in phylogenetics, by integrating large-scale phylogenomic analyses with the search for punctuate, defining characters.

Lastly, the integration of artificial intelligence (AI) opens up vast possibilities to improve our understanding of the ToL through multiple improvement axes (Fig. 2). It is likely that AI tools will become indispensable in managing and interpreting the vast amounts of genomic data generated by modern sequencing technologies (as an example, the size of GenBank increased more than 16-fold over the last 10 years). AI algorithms excel at identifying patterns in large data sets. In the context of genomics, AI can detect similarities and differences in sequences of different organisms, which will improve the quality of the metagenomic binning and generally the detection of sequence contamination. From the perspective of evolutionary bioinformatics, AI can improve the accuracy and efficiency of sequence alignment and phylogenetic reconstruction, handling the complexity of data sets comprising thousands to millions of sequences. AI is also adept at detecting outliers or anomalies in data. This can help identify genes or sequences that do not follow the expected evolutionary patterns, possibly indicating events like horizontal gene transfer, gene duplication, or convergent evolution (Mo et al. 2023; van Hooff and Eme 2023). It also opens up the possibility of using protein structures instead of sequences to infer difficult relationships (Moi et al. 2023). Finally, AI can integrate and analyze data from different genomic sources (e.g. whole-genome sequencing, transcriptomics, and proteomics), which could provide a more holistic view of an organism's evolution. By correlating changes across different levels of biological information, AI can help uncover the multifaceted relationships between genotype and phenotype, as well as between different species and their environments. By analyzing genomic sequences of a wide array of species, AI models could predict how certain genes contribute to specific traits or behaviors, shedding light on the evolutionary pressures that have shaped current life forms. Through these and other applications, AI will not only help tackling the challenge of managing and analyzing enormous quantities of genomic data but also will enhance our ability to decipher the complex web of evolutionary relationships that connect all living organisms.

The synergy of these developments—technological advances in sampling and sequencing, methodological developments in phylogenetics, and computational innovation leading to more efficient, robust, and tractable algorithms—heralds an exciting era in unraveling the complexities of the ToL, bringing us closer to a comprehensive understanding of life's evolutionary tapestry.

Acknowledgments

We sincerely thank the three anonymous reviewers for their constructive comments. We apologize to our colleagues

whose work was not adequately acknowledged due to citation limitations.

Funding

L.E. was supported by grants from the European Research Council (ERC Starting grant 803151), the Moore-Simons Project Call on the Origin of the Eukaryotic Cell (Simons Foundation 812811), and the ANR DArchFolds (ANR-22-CE02-0012-02).

Data Availability

There are no new data associated with this work.

Literature Cited

Al Jewari C, Baldauf SL. An excavate root for the eukaryote Tree of Life. Sci Adv. 2023:9(17):eade4973. https://doi.org/10.1126/sciadv.ade4973.

Andersson JO. Gene transfer and diversification of microbial eukaryotes. Annu Rev Microbiol. 2009:63(1):177–193. https://doi.org/10.1146/annurev.micro.091208.073203.

Baker BA, Gutiérrez-Preciado A, Rodriguez del Rio A, McCarthy C, López-García P, Huerta-Cepas J, Susko E, Roger AJ, Eme L, Moreira D. Several independent adaptations of archaea to hypersaline environments. Nat Microbiol. 2024:1–12. https://doi.org/10.1038/s41564-024-01647-4.

Balaban M, Jiang Y, Zhu Q, McDonald D, Knight R, Mirarab S. Generation of accurate, expandable phylogenomic trees with uDance. Nat Biotechnol. 2023:1–10. https://doi.org/10.1038/s41587-023-01868-8.

Betts HC, Puttick MN, Clark JW, Williams TA, Donoghue PCJ, Pisani D. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. Nat Ecol Evol. 2018:2(10):1556–1562. https://doi.org/10.1038/s41559-018-0644-x.

Boscaro V, Manassero V, Keeling PJ, Vannini C. Single-cell microbiomics unveils distribution and patterns of microbial symbioses in the natural environment. Microb Ecol. 2023:85(1):307–316. https://doi.org/10.1007/s00248-021-01938-x.

Brinkmann H, Philippe H. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol. 1999:16(6):817–825. https://doi.org/10.1093/oxfordjournals.molbev.a026166.

Brown JR, Doolittle WF. Root of the universal Tree of Life based on ancient aminoacyl-tRNA synthetase gene duplications. Proc Natl Acad Sci U S A. 1995:92(7):2441–2445. https://doi.org/10.1073/pnas.92.7.2441.

Brown MW, Heiss AA, Kamikawa R, Inagaki Y, Yabuki A, Tice AK, Shiratori T, Ishida K-I, Hashimoto T, Simpson AGB, et al. Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. Genome Biol Evol. 2018:10(2):427–433. https://doi.org/10.1093/gbe/evy014.

Burki F, Roger AJ, Brown MW, Simpson AGB. The new tree of eukaryotes. Trends Ecol Evol. 2020:35(1):43–55. https://doi.org/10.1016/j.tree.2019.08.008.

Cerón-Romero MA, Fonseca MM, de Oliveira Martins L, Posada D, Katz LA. Phylogenomic analyses of 2,786 genes in 158 lineages support a root of the eukaryotic Tree of Life between opisthokonts and all other lineages. Genome Biol Evol. 2022:14(8):evac119. https://doi.org/10.1093/gbe/evac119.

Coleman GA, Davín AA, Mahendrarajah TA, Szánthó LL, Spang A, Hugenholtz P, Szöllősi GJ, Williams TA. A rooted phylogeny resolves early bacterial evolution. Science. 2021:372(6542):eabe0511. https://doi.org/10.1126/science.abe0511.

Comte A, Tricou T, Tannier E, Joseph J, Siberchicot A, Penel S, Allio R, Delsuc F, Dray S, de Vienne DM. 2023. Phylter: efficient identification of outlier sequences in phylogenomic datasets. bioRxiv. 2023.02.02.526888. https://doi.org/10.1101/2023.02.02.526888.

Dagan T, Roettger M, Bryant D, Martin W. Genome networks root the Tree of Life between prokaryotic domains. Genome Biol Evol. 2010:2:379–392. https://doi.org/10.1093/gbe/evq025.

De Farias ST, Bonato MCM. Preferred codons and amino acid couples in hyperthermophiles. Genome Biol. 2002:3(8):PREPRINT0006. https://doi.org/10.1186/gb-2002-3-8-preprint0006.

Del Amparo R, Arenas M. Consequences of substitution model selection on protein ancestral sequence reconstruction. Mol Biol Evol. 2022:39(7). https://doi.org/10.1093/molbev/msac144.

Derelle R, Lang BF. Rooting the eukaryotic tree with mitochondrial and bacterial proteins. Mol Biol Evol. 2012:29(4):1277–1289. https://doi.org/10.1093/molbev/msr295.

Dia A, Cheeseman IH. Single-cell genome sequencing of protozoan parasites. Trends Parasitol. 2021:37(9):803–814. https://doi.org/10.1016/j.pt.2021.05.013.

Eme L, Sharpe SC, Brown MW, Roger AJ. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. Cold Spring Harb Perspect Biol. 2014:6(8):3–4. https://doi.org/10.1101/cshperspect.a016139.

Eme L, Tamarit D, Caceres EF, Stairs CW, De Anda V, Schön ME, Seitz KW, Dombrowski N, Lewis WH, Homa F, et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. Nature. 2023:618(7967):992–999. https://doi.org/10.1038/s41586-023-06186-2.

Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. Syst Biol. 1978:27(4):401–410. https://doi.org/10.1093/sysbio/27.4.401.

Fournier GP, Gogarten JP. Rooting the ribosomal Tree of Life. Mol Biol Evol. 2010:27(8):1792–1801. https://doi.org/10.1093/molbev/msq057.

Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. Unique amino acid composition of proteins in halophilic bacteria. J Mol Biol. 2003:327(2):347–357. https://doi.org/10.1016/S0022-2836(03)00150-5.

Galindo LJ, Torruella G, Moreira D, Eglit Y, Simpson AGB, Völcker E, Clauß S, López-García P. Combined cultivation and single-cell approaches to the phylogenomics of *Nucleariid amoebae*, close relatives of fungi. Philos Trans R Soc Lond B Biol Sci. 2019:374(1786):20190094. https://doi.org/10.1098/rstb.2019.0094.

Gawryluk RMR, Tikhonenkov DV, Hehenberger E, Husnik F, Mylnikov AP, Keeling PJ. Non-photosynthetic predators are sister to red algae. Nature. 2019:572(7768):240–243. https://doi.org/10.1038/s41586-019-1398-6.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, et al. Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci U S A. 1989:86(17):6661–6665. https://doi.org/10.1073/pnas.86.17.6661.

Grossart H-P, Massana R, McMahon KD, Walsh DA. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. Limnol Oceanogr. 2020:65(S1):S2–S20. https://doi.org/10.1002/lno.11382.

He D, Fiz-Palacios O, Fu C-J, Fehling J, Tsai C-C, Baldauf SL. An alternative root for the eukaryote Tree of Life. Curr Biol. 2014:24(4):465–470. https://doi.org/10.1016/j.cub.2014.01.036.

Hernández-Plaza A, Szklarczyk D, Botas J, Cantalapiedra CP, Giner-Lamia J, Mende DR, Kirsch R, Rattei T, Letunic I, Jensen LJ,

et al. eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucleic Acids Res. 2023:51(D1):D389–D394. https://doi.org/10.1093/nar/gkac1022.
Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. A new view of the Tree of Life. Nat Microbiol. 2016:1(5):16048. https://doi.org/10.1038/nmicrobiol.2016.48.
Hugenholtz P, Tyson GW. Microbiology: metagenomics. Nature. 2008:455(7212):481–483. https://doi.org/10.1038/455481a.
Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, Takano Y, Uematsu K, Ikuta T, Ito M, et al. Isolation of an archaeon at the prokaryote–eukaryote interface. Nature. 2020:577(7791):519–525. https://doi.org/10.1038/s41586-019-1916-6.
Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci U S A. 1989:86(23):9355–9359. https://doi.org/10.1073/pnas.86.23.9355.
Kapli P, Flouri T, Telford MJ. Systematic errors in phylogenetic trees. Curr Biol. 2021:31(2):R59–R64. https://doi.org/10.1016/j.cub.2020.11.043.
Lake JA. Evolving ribosome structure: domains in archaebacteria, eubacteria, eocytes and eukaryotes. Annu Rev Biochem. 1985:54(1):507–530. https://doi.org/10.1146/annurev.bi.54.070185.002451.
Lax G, Eglit Y, Eme L, Bertrand EM, Roger AJ, Simpson AGB. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. Nature. 2018:564(7736):410–414. https://doi.org/10.1038/s41586-018-0708-8.
López-García P, Moreira D. Tracking microbial biodiversity through molecular and genomic ecology. Res Microbiol. 2008:159(1):67–73. https://doi.org/10.1016/j.resmic.2007.11.019.
Lopez P, Forterre P, Philippe H. The root of the Tree of Life in the light of the covarion model. J Mol Evol. 1999:49(4):496–508. https://doi.org/10.1007/PL00006572.
Mahendrarajah TA, Moody ERR, Schrempf D, Szánthó LL, Dombrowski N, Davín AA, Pisani D, Donoghue PCJ, Szöllősi GJ, Williams TA, et al. ATP synthase evolution on a cross-braced dated Tree of Life. Nat Commun. 2023:14(1):7456. https://doi.org/10.1038/s41467-023-42924-w.
Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJG. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature. 2018:557(7703):101–105. https://doi.org/10.1038/s41586-018-0059-5.
Mo Y, Hahn M, Smith M. Applications of machine learning in phylogenetics. EcoEvoRxiv. 2023. https://doi.org/10.32942/x2xg7g.
Moi D, Bernard C, Steinegger M, Nevers Y, Langleib M, Dessimoz C. 2023. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. bioRxiv. 2023.09.19.558401. https://doi.org/10.1101/2023.09.19.558401.
Muñoz-Gómez SA, Susko E, Williamson K, Eme L, Slamovits CH, Moreira D, López-García P, Roger AJ. Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. Nat Ecol Evol. 2022:6(3):253–262. https://doi.org/10.1038/s41559-021-01638-2.
Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 2022:50(D1):D785–D794. https://doi.org/10.1093/nar/gkab776.
Philippe H, de Vienne DM, Ranwez V, Roure B, Baurain D, Delsuc F. Pitfalls in supermatrix phylogenomics. Eur J Taxon. 2017:283:1–25.
Pittis AA, Gabaldón T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. Nature. 2016:531(7592):101–104. https://doi.org/10.1038/nature16941.
Reed CJ, Lewis H, Trejo E, Winston V, Evilia C. Protein adaptations in archaeal extremophiles. Archaea. 2013:2013:373275.
Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, Wideman JG, Burki F, de Vargas C. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. Peer Community J. 2022:2. https://doi.org/10.24072/pcjournal.173.
Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, Darling A, Malfatti S, Swan BK, Gies EA, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013:499(7459):431–437. https://doi.org/10.1038/nature12352.
Rodrigues-Oliveira T, Wollweber F, Ponce-Toledo RI, Xu J, Rittmann SKMR, Klingl A, Pilhofer M, Schleper C. Actin cytoskeleton and complex cell architecture in an Asgard archaeon. Nature. 2022:613(7943):332–339. https://doi.org/10.1038/s41586-022-05550-y.
Ruiz-Trillo I, Kin K, Casacuberta E. The origin of metazoan multicellularity: a potential microbial black swan event. Annu Rev Microbiol. 2023:77(1):499–516. https://doi.org/10.1146/annurev-micro-032421-120023.
Schön ME, Zlatogursky VV, Singh RP, Poirier C, Wilken S, Mathur V, Strassert JFH., Pinhassi J, Worden AZ, Keeling PJ, et al. Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. Nat Commun. 2021:12(1):6651. https://doi.org/10.1038/s41467-021-26918-0.
Sibbald SJ, Eme L, Archibald JM, Roger AJ. Lateral gene transfer mechanisms and pan-genomes in eukaryotes. Trends Parasitol. 2020:36(11):927–941. https://doi.org/10.1016/j.pt.2020.07.014.
Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015:521(7551):173–179. https://doi.org/10.1038/nature14447.
Steenwyk JL, Li Y, Zhou X, Shen X-X, Rokas A. Incongruence in the phylogenomics era. Nat Rev Genet. 2023:24(12):834–850. https://doi.org/10.1038/s41576-023-00620-x.
Susko E, Lincker L, Roger AJ. Accelerated estimation of frequency classes in site-heterogeneous profile mixture models. Mol Biol Evol. 2018:35(5):1266–1283. https://doi.org/10.1093/molbev/msy026.
Susko E, Roger AJ. On reduced amino acid alphabets for phylogenetic inference. Mol Biol Evol. 2007:24(9):2139–2150. https://doi.org/10.1093/molbev/msm144.
Susko E, Roger AJ. On the use of information criteria for model selection in phylogenetics. Mol Biol Evol. 2020:37(2):549–562. https://doi.org/10.1093/molbev/msz228.
Susko E, Roger AJ. Long branch attraction biases in phylogenetics. Syst Biol. 2021:70(4):838–843. https://doi.org/10.1093/sysbio/syab001.
Susko E, Steel M, Roger AJ. Conditions under which distributions of edge length ratios on phylogenetic trees can be used to order evolutionary events. J Theor Biol. 2021:526:110788. https://doi.org/10.1016/j.jtbi.2021.110788.
Tice AK, Žihala D, Pánek T, Jones RE, Salomaki ED, Nenarokov S, Burki F, Eliáš M, Eme L, Roger AJ, et al. PhyloFisher: a phylogenomic package for resolving eukaryotic relationships. PLoS Biol. 2021:19(8):e3001365. https://doi.org/10.1371/journal.pbio.3001365.
Tikhonenkov DV, Mikhailov KV, Gawryluk RMR, Belyaev AO, Mathur V, Karpov SA, Zagumyonnyi DG, Borodina AS, Prokina KI, Mylnikov AP, et al. Microbial predators form a new supergroup of eukaryotes. Nature. 2022:612(7941):714–719. https://doi.org/10.1038/s41586-022-05511-5.
van Hooff JJE, Eme L. 2023. Lateral gene transfer leaves lasting traces in Rhizaria. bioRxiv. 2023.01.27.525846. https://doi.org/10.1101/2023.01.27.525846.
Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004:304(5667):66–74. https://doi.org/10.1126/science.1093857.

Wideman JG, Monier A, Rodríguez-Martínez R, Leonard G, Cook E, Poirier C, Maguire F, Milner DS, Irwin NAT, Moore K, et al. Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. Nat Microbiol. 2020:5(1):154–165. https://doi.org/10.1038/s41564-019-0605-4.

Williams TA, Davín AA, Morel B, Szánthó LL, Spang A, Stamatakis A, Hugenholtz P, Szöllősi GJ. Parameter estimation and species tree rooting using ALE and GeneRax. Genome Biol Evol. 2023:15(7). https://doi.org/10.1093/gbe/evad134.

Williams TA, Embley TM. Archaeal 'dark matter' and the origin of eukaryotes. Genome Biol Evol. 2014:6(3):474–481. https://doi.org/10.1093/gbe/evu031.

Williams TA, Schrempf D, Szöllősi GJ, Cox CJ, Foster PG, Embley TM. Inferring the deep past from molecular data. Genome Biol Evol. 2021:13(5). https://doi.org/10.1093/gbe/evab067.

Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJG, Embley TM. Integrative modeling of gene and genome evolution roots the archaeal Tree of Life. Proc Natl Acad Sci U S A. 2017:114(23):E4602–E4611. https://doi.org/10.1073/pnas.1618463114.

Zhaxybayeva O, Doolittle WF. Lateral gene transfer. Curr Biol. 2011:21(7):R242–R246. https://doi.org/10.1016/j.cub.2011.01.045.

Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, Belda-Ferre P, Al-Ghalith GA, Kopylova E, McDonald D, et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. Nat Commun. 2019:10(1):5477. https://doi.org/10.1038/s41467-019-13443-4.

Associate editor: John Archibald