PLOS BIOLOGY

CONSENSUS VIEW

Reconstructing the last common ancestor of all eukaryotes

Thomas A. Richards[1]*, Laura Eme[2,3]*, John M. Archibald[4]*, Guy Leonard[1], Susana M. Coelho[5], Alex de Mendoza[6], Christophe Dessimoz[7,8], Pavel Dolezal[9], Lillian K. Fritz-Laylin[10], Toni Gabaldón[11,12,13,14], Vladimír Hampl[9], Geert J. P. L. Kops[15], Michelle M. Leger[16,17], Purificacion Lopez-Garcia[2], James O. McInerney[18], David Moreira[2], Sergio A. Muñoz-Gómez[19], Daniel J. Richter[16], Iñaki Ruiz-Trillo[13,16], Alyson E. Santoro[20], Arnau Sebé-Pedrós[12,21], Berend Snel[22], Courtney W. Stairs[23], Eelco C. Tromer[24], Jolien J. E. van Hooff[25], Bill Wickstead[26], Tom A. Williams[27], Andrew J. Roger[4]*, Joel B. Dacks[28,29,30]*, Jeremy G. Wideman[31]*

1 Department of Biology, University of Oxford, Oxford, United Kingdom, **2** Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Gif-sur-Yvette, France, **3** Department of Cell & Molecular Biology, The University of Rhode Island, Kingston, Rhode Island, United States of America, **4** Department of Biochemistry and Molecular Biology and the Institute for Comparative Genomics, Dalhousie University, Halifax, Canada, **5** Department of Algal Development and Evolution, Max Planck Institute for Biology Tübingen, Tübingen, Germany, **6** School of Biological and Behavioural Sciences, Queen Mary University of London, London, United States of America, **7** Department of Computational Biology, University of Lausanne, Lausanne, Switzerland, **8** Swiss Institute of Bioinformatics, Lausanne, Switzerland, **9** Charles University, Faculty of Science, Department of Parasitology, BIOCEV, Vestec, Czech Republic, **10** Department of Biology, University of Massachusetts Amherst, Amherst, Massachusetts, United States of America, **11** Barcelona Supercomputing Centre (BSC-CNS), Barcelona, Spain, **12** Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain, **13** Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain, **14** CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain, **15** Hubrecht Institute-KNAW, Oncode Institute, UMC Utrecht, Utrecht, the Netherlands, **16** Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Spain, **17** Okinawa Institute of Science and Technology Graduate University (OIST), Okinawa, Japan, **18** Department of Evolution, Ecology and Behaviour, Institute of Infection, Veterinary and Ecological Sciences, University of Liverpool, Liverpool, United Kingdom, **19** Department of Biological Sciences, Purdue University, West Lafayette, Indiana, United States of America, **20** Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California, United States of America, **21** Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, **22** Theoretical Biology and Bioinformatics, Department of Biology, Faculty of Science, Utrecht University, Utrecht, the Netherlands, **23** Department of Biology, Lund University, Lund, Sweden, **24** Cell Biochemistry, Groningen Biomolecular Sciences and Biotechnology Institute, Rijksuniversiteit Groningen, Groningen, the Netherlands, **25** Laboratory of Microbiology, Wageningen University & Research, Wageningen, the Netherlands, **26** School of Life Sciences, University of Nottingham, Nottingham, United Kingdom, **27** School of Biological Sciences, University of Bristol, Bristol, United Kingdom, **28** Division of Infectious Diseases, Department of Medicine, and Department of Biological Sciences, University of Alberta, Edmonton, Canada, **29** Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic, **30** Centre for Life's Origins and Evolution, Department of Genetics, Evolution, & Environment, University College, London, United Kingdom, **31** Center for Mechanisms of Evolution, School of Life Sciences, Arizona State University, Tempe, Arizona, United States of America

* thomas.richards@biology.ox.ac.uk (TAR); laura.eme@universite-paris-saclay.fr (LE); jmarchib@dal.ca (JMA); aroger@dal.ca (AJR); dacks@ualberta.ca (JBD); Jeremy.Wideman@asu.edu (JGW)

Citation: Richards TA, Eme L, Archibald JM, Leonard G, Coelho SM, de Mendoza A, et al. (2024) Reconstructing the last common ancestor of all eukaryotes. PLoS Biol 22(11): e3002917. https://doi.org/10.1371/journal.pbio.3002917

Published: November 25, 2024

Funding: This work was supported by the Royal Society University Research Fellowship grant URF \R\191005 to TAR; Berlin Institute for Advanced Study (Wissenschaftskolleg zu Berlin) Fellowship to TAR; Gordon and Betty Moore Foundation Grant (GBMF9730) to TAR; HORIZON EUROPE European Research Council grant (803151) to LE; Simons Foundation Grant (735923LPI) to LE and AJR; Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2019-05058) to JMA; Gordon and Betty Moore Foundation Grant (GBMF5782) to JMA; Arthur B. McDonald Chair of Research Excellence at Dalhousie University to JMA; Natural Sciences and Engineering Research Council of Canada (RGPIN-2022-05430) to AJR; Natural Sciences and Engineering Research Council of Canada (RES0043758 & RES0046091) to JBD; National Science Foundation (2119963 & 2405455) to

Abstract

Understanding the origin of eukaryotic cells is one of the most difficult problems in all of biology. A key challenge relevant to the question of eukaryogenesis is reconstructing the gene repertoire of the last eukaryotic common ancestor (LECA). As data sets grow, sketching an accurate genomics-informed picture of early eukaryotic cellular complexity requires

JGW, and Gordon and Betty Moore Foundation Grant (GBMF10600) to JGW. The funders had no role in the decision to publish, or the preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: ESP, eukaryote signature protein; FECA, First Eukaryotic Common Ancestor; HGT, horizontal gene transfer; HMM, hidden Markov model; LECA, last eukaryotic common ancestor.

provision of analytical resources and a commitment to data sharing. Here, we summarise progress towards understanding the biology of LECA and outline a community approach to inferring its wider gene repertoire. Once assembled, a robust LECA gene set will be a useful tool for evaluating alternative hypotheses about the origin of eukaryotes and understanding the evolution of traits in all descendant lineages, with relevance in diverse fields such as cell biology, microbial ecology, biotechnology, agriculture, and medicine. In this Consensus View, we put forth the status quo and an agreed path forward to reconstruct LECA's gene content.

Introduction

The origin of the eukaryotic cell is one of the most significant evolutionary transitions in the history of life [1]. Eukaryotes are fundamentally different from their prokaryotic relatives (Bacteria and Archaea) in how the cell is organised, how these cells "feed," move, and respond to stimuli, and how their genes are structured and expressed. Eukaryogenesis is a subject of active research and debate [2–9]). Because the eukaryotic cell evolved between 1.5 and 2.5 billion years ago [10–12], direct experimental approaches are limited and phylogenetic analyses are vulnerable to methodological artefacts [13–15]. These are problems compounded by having no other major transition of a similar age and complexity to which eukaryogenesis can be compared. Consensus on how eukaryotes first arose is thus lacking, and it is unclear how best to approach unanswered questions in order to maximise the effectiveness of future research.

Debates about eukaryogenesis span multiple disciplines including microbiology, paleobiology, and cell biology; yet they often rely heavily on phylogenomic investigations [16]. These analyses involve inferring the distribution and evolutionary history of gene families across eukaryotic and prokaryotic diversity. Here, we provide recommendations for establishing a robust phylogenomics-based picture of the genetic, metabolic, and cellular repertoires of the ancestral form(s) that gave rise to all extant eukaryotes, i.e., the last eukaryotic common ancestor (LECA) [17,18]. The goal is to produce a resolved picture of LECA and a tractable gene repertoire. The latter will serve as an important data set for understanding the prokaryotic origin-(s) of the eukaryotes and to compare different hypotheses pertinent to early eukaryotic cell evolution.

A minimal consensus on the origin of eukaryotes

Most researchers accept that LECA originated after an association of at least 2 organisms descending from prokaryotes of evolutionarily distinct lineages—one arising from within the Archaea [19,20], likely within the Asgardarchaeota [8,21], and the other related to Alphaproteobacteria [21–24]. We refer to this scenario as the "two+" model of eukaryogenesis (i.e., 2 partners coupled with significant evolutionary change in the fundamental cell biology of this emerging form). This baseline scenario provides a starting point for comparing alternative hypotheses. For example, many variant hypotheses suggest that additional lineages contributed to eukaryogenesis [25], e.g., a "third partner" arising from deltaproteobacteria [26] or chlamydia-like bacteria [27], while others have suggested an alternative starting point to eukaryogenesis from close to the planctomycetes [28]. Still others have suggested that viruses were major contributors [29–33], although viruses of various forms certainly acted as agents moving genes between lineages throughout the history of eukaryotic evolution [30,34], thus making it difficult to identify early viral contributions to eukaryogenesis. Which auxiliary lineages participated, when and how—either by bursts of horizontal gene transfer (HGT) from short-lived microbial associations, or longer-term integrations like the endosymbiotic processes that led to the mitochondrion or the plastid [22,35]—are long-standing questions in the eukaryogenesis debate.

Although there is broad acceptance that living eukaryotes arose from a common ancestor that had genetic and cellular features of mixed archaeal and bacterial ancestry, hypotheses differ as to which cellular lineage is proposed to have "encapsulated" the other; some suggest an archaeon took up a bacterium [36–40] while others argue the opposite [26]. Other models envisage the alphaproteobacterium-related mitochondrion as having been established via phagocytosis by a proto-eukaryote of archaeal ancestry that already possessed many of the canonical features of extant eukaryotes, such as a cytoskeleton, endomembrane system, and nucleus [41]. These various models have been discussed extensively (e.g., [21,26,36–40,42,43]) with little resolution.

Several contributions have sought a clear definition of terms relating to eukaryogenesis to help frame wider debate [21,44–46]. The First Eukaryotic Common Ancestor (FECA) can be defined as the first descendant—on the eukaryotic side—of the last common ancestor of an Asgardarchaeota lineage and the eukaryotes [44,46] (i.e., the first organism whose living descendants only include eukaryotes and no other extant lines). Under the two+ model at least one other FECA lineage can be said to have existed, i.e., the first descendant of the last common ancestor of the alphaproteobacteria-related progenitor and the eukaryotes [23,24,44,46]. To simplify discussion, we refer to this latter FECA as the first mitochondrial common ancestor or FMCA (pronounced "Firmca") [21]. There could be additional FECAs if a third or even fourth lineage were also involved in eukaryogenesis as suggested by some analyses [47]. The divergences of eukaryotes from Asgard archaea and from Alphaproteobacteria are important because they mark the beginning of the period in which the hallmark features of eukaryotes might have evolved. However, crucially and perhaps counterintuitively, there is no implication that archaeal FECA or FMCA were more eukaryote-like than their immediate prokaryotic ancestors, because the cellular features we now associate with eukaryotes might have evolved at any point on the stems between either the archaeal FECA and LECA or between FMCA and LECA [46].

At present, the unresolved gap between the archaeal FECA and LECA, and indeed FMCA and LECA, makes it difficult to infer the order and nature of events between these ancestral forms [21]. Additional sampling of lineages that branch closer to the eukaryotes than currently known prokaryotes would add greater resolution in understanding eukaryogenesis. Attempts have been made to reconstruct the order of prokaryotic gene acquisition (e.g., Asgard, alphaproteobacterial, or additional prokaryotic contributions) between these 2 points [47,48], but our understanding of this process remains limited. Analyses of shared gene content between Asgardarchaeota and extant eukaryotes have been useful in gaining a clearer picture of one set of contributions to LECA [8,9,49]. However, reconstructing the contribution of any FECA—including FMCA—depends on knowing the gene content of LECA.

How can reconstruction of LECA inform our understanding of eukaryogenesis?

In order to appropriately understand LECA, 2 related problems need to be addressed:

i. What was the molecular cell biology of LECA? Specifically, what molecular components and cellular systems evolved prior to LECA? Which of those systems arose later, as the eukaryotic lineages diverged?

ii. Where did LECA come from? Specifically, which prokaryotic subgroups were the key partners and which genes did they contribute? Conversely, which genes evolved de novo during the FECA-to-LECA transition/s?

If we can achieve consensus on these points, understanding LECA would enable us to define the endpoint of eukaryogenesis. This would be the end state at which all eukaryogenesis models must arrive and a starting point for understanding the evolution of the major

eukaryotic groups and the cellular systems that arose within them (i.e., a baseline comparator for polarising all subsequent evolutionary transitions).

A consensus LECA gene repertoire also provides a framework for judging the relative merits of different eukaryogenesis models. Specifically, we can use these data to determine if "eukaryogenesis model X" has merit (utility) because it is consistent with the inferred evolutionary histories of the genes present in the LECA gene repertoire. For example, if a pattern of "third-party" ancestry (e.g., deltaproteobacteria or chlamydia [26,27,43,50,51]) is identified in a significant proportion of LECA gene trees (e.g., **Fig 1A**), then a three-partner eukaryogenesis model could then be favoured. We note that without evidence of an endosymbiotically derived compartment or genome, it would not be possible to distinguish between bursts of gene transfer from transient microbial associations, or a longer-term integration similar to the process which generated the endosymbiotically derived organelles. However, such patterns may theoretically be distinguishable from serial HGT processes as identified, for example, from viral contribution (e.g., [34]) using phylostratigraphy-like approaches [52]. However, if a substantial third-party prokaryotic signal is absent (e.g., **Fig 1B**), phylogenetic patterns provide little support beyond the two+ model.

Fig 1 compares a range of possible outcomes from LECA analyses, not just the presence or absence of a third-party contributor. For example, a relatively large LECA gene repertoire (Fig 1A) versus a smaller one (Fig 1B), implies a very different relative role for gene family gain and expansion post LECA. Furthermore, the models demonstrate very different roles for de novo gene evolution and the relative contribution of prokaryotic genes. For simplicity, these factors are shown in 2 distinct constellations. This is not to say that these are the only constellations possible—indeed different combinations of the characteristics illustrated across the 2 models can be imagined. This is not to trivialise the problem or the complexity of the data; there is a range of possible outcomes, and LECA reconstructions may identify a result somewhere between the 2 extremes shown in **Fig 1**. Our goal is to outline how different models might be supported, refuted, and appropriately modified in response to data, thereby minimising polarised debates about what genes, molecular systems, and cellular processes were—and were not—"important" for eukaryogenesis. A community-wide effort to define LECA will permit informed comparisons of different models so that they can be judged on their relative merits.

Understanding the mixed ancestry of LECA

Of the fraction of genes present in LECA that possess obvious prokaryotic homology, only a small fraction can be definitively shown to be of alphaproteobacterial or asgardarchaeal origin. In a recent study of gene family evolution in eukaryotes [48], 10,233 Pfam domain families were inferred to be present in LECA. Of these, 4,335 families were acquired from prokaryotic sources, and 77% of these acquisitions were identified as having bacterial ancestry; 7% appeared to be of alphaproteobacterial-like origin. Approximately 16% of the prokaryotic acquisitions were identified as "archaeal" with only 7% specifically of Asgardarchaeota ancestry [48]. However, raw percentages do not necessarily linearly correlate with evolutionary importance. Few gene acquisitions can give rise to fundamental systems; consequently, comparisons using such statistics have to be considered carefully. Nonetheless, such data have profound implications for the two+ basic model and suggest a LECA model more closely aligned with Fig 1B (a scenario in which the ancestry of most prokaryotic genes cannot be traced back to specific donors, e.g., the Asgardarchaeota or the Alphaproteobacteria) rather than Fig 1A. What might this mean for eukaryogenesis?

The large number of LECA genes that do not trace back to either Asgardarchaeota or Alphaproteobacteria has been interpreted as evidence for additional or alternative prokaryotic or viral contributors to LECA (e.g., [25–31]). However, the presence of additional genomes and/or

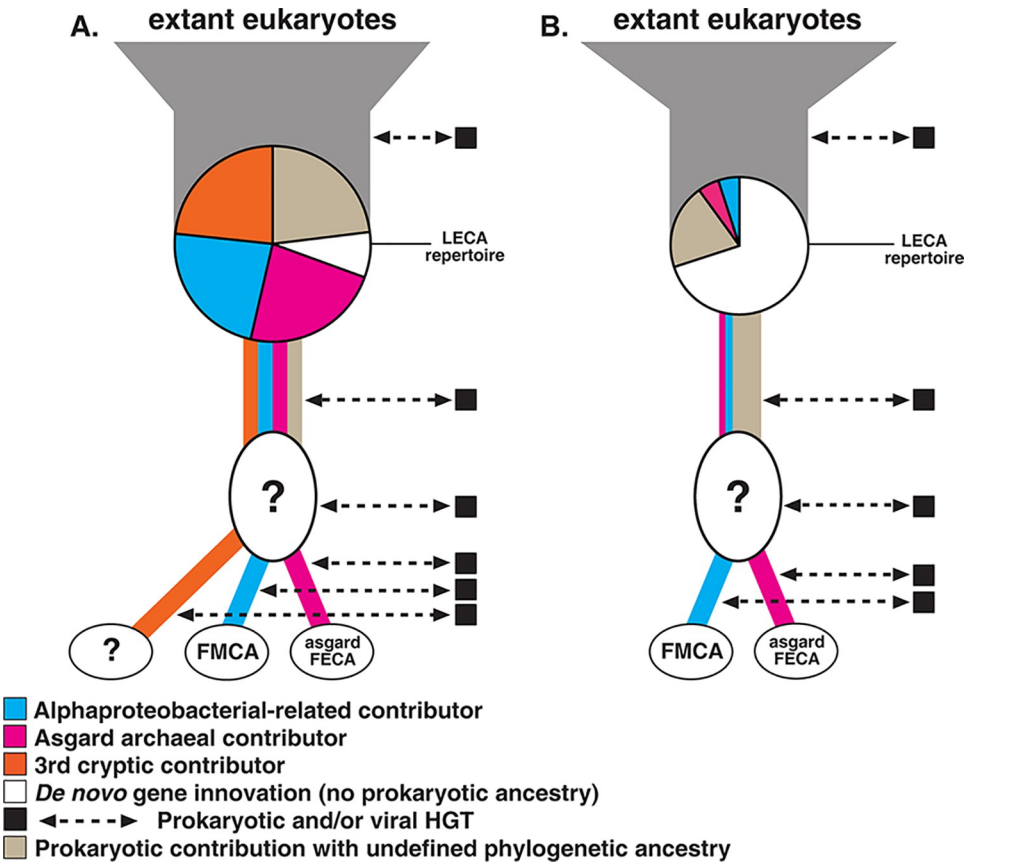

Fig 1. Genetic contributions to LECA. LECA's gene repertoire was chimeric, containing genes derived from the Asgardarchaeota-derived host cell, mitochondrial endosymbiont, and potentially other prokaryotic sources, along with a set of eukaryote-specific genes that evolved after the divergence of eukaryotes from prokaryotes. The number of sources, and the proportions and identities of genes from each source, remain uncertain but can be investigated using the approach articulated in the main text of this paper. Here, we illustrate 2 possible LECA reconstructions that are broadly compatible with what is currently known about eukaryotic gene origins. **(A)** Shows a larger LECA gene repertoire reconstruction as indicated by the large pie chart. Such an inference may be the result of relatively few gene innovations post LECA, as indicated by the modest expansion after LECA leading to extant eukaryotic diversity. This hypothetical model also shows strong Asgardarchaeota and alphaproteobacterial signals and a strong additional signal from a "third party" contributor. This "third signal" could be used to argue for the role of 3 contributing lineages to eukaryogenesis beyond the two+ model. Here, the fraction of genes of de novo gene evolution (i.e., bona fide ESPs) is relatively small. The proportion of gene families of prokaryotic ancestry with poor phylogenetic resolution is not a dominant ancestral signal. **(B)** Shows a smaller LECA gene repertoire reconstruction as indicated by a smaller pie chart. Such an inference may indicate a larger-scale gene innovation post LECA, as indicated by the wider expansion after the LECA lineage leading to extant eukaryotic diversity. In this hypothetical model, the LECA repertoire with identifiable prokaryotic origin is dominated by genes of undefined ancestry. This model also shows that the LECA gene families of de novo gene ancestry (ESPs) is extensive. Only a tiny proportion of gene families present in LECA can be accurately attributed to either the Asgardarchaeota or the Alphaproteobacteria. The question marks inside the ovals on both models **A** and **B** indicate an unknown order of contribution and/or unknown contributing lineages. Dashed double arrow-headed lines indicate possible HGT contributions throughout eukaryogenesis and subsequent diversification of eukaryotes. Not all aspects of these models are mutually exclusive; for example, a large LECA repertoire (as shown in A) could be combined with a two+ model for ancestry (as shown in B). ESP, eukaryote signature protein; HGT, horizontal gene transfer; LECA, last eukaryotic common ancestor.

https://doi.org/10.1371/journal.pbio.3002917.g001

compartments within the eukaryotic cell, separate from the nucleus, and hosting these genes (like mitochondria and plastids), would provide indisputable evidence for additional prokaryotic partners. In the absence of such evidence, an alternative explanation is that early eukaryotic forms engaged in HGT [114] both into and out of the FECA-to-LECA lineages, a pattern seen in extant eukaryotes [115–119]. An additional variant of the HGT explanation is that these detected

prokaryotic ancestries are footprints of prior transient endosymbiotic associations that laid the groundwork for the eventual mitochondrial endosymbiosis, as seen in more recent symbiotic associations and organelle acquisition events [120,121]. Another, not mutually exclusive, explanation is the "fluid prokaryotic chromosome model," which posits that HGT between prokaryotes has been so frequent and ongoing that the genomes of the 2 prokaryotic lineages constituting the two + model were themselves highly mosaic at the time of eukaryogenesis. More generally, incomplete taxon sampling and/or complex patterns of gene retention and loss since eukaryogenesis likely contributed to the mixed prokaryotic phylogenetic affinities seen in extant eukaryotes [122,123].

Many genes inferred to have been present in LECA do not currently have identifiable prokaryotic homologs (e.g., [48,63–65,124]). Such genes encode possible "eukaryotic signature proteins" or ESPs [39,124–127]. For example, in a 2021 study by Vosseberg and colleagues [48], 58% of the eukaryotic gene families analysed had no identifiable prokaryotic ancestry, a number that is likely to be further revised as methods change and more prokaryotes (and eukaryotes) are sampled. This reinforces the view that eukaryogenesis was a radical transition that triggered—and indeed was to a certain extent enabled by—gene family expansion. However, the discovery that Asgardarchaeota possess a subset of the genes previously classified as ESPs has somewhat altered this picture [8,9,21,49]. Nonetheless, numerous proteins not yet found in the Asgardarchaeota remain as candidate ESPs. So where did the significant proportion of LECA genes with no apparent similarity to prokaryotic genes come from? Beyond de novo gene evolution (i.e., new genes arising from non-coding DNA), it is possible that an unsampled (or extinct) third-party "prokaryotic" donor group possesses (or possessed) genes uniquely shared with the eukaryotes. It is also likely that a high rate of sequence evolution at eukaryogenesis currently prevents us from identifying the prokaryotic homologs of many ESPs based on sequence similarity alone.

A final consideration when trying to understand the ancestry of the genetic constituents of LECA is the limitations of current phylogenetic methods. Even the best methods currently available may struggle to model sequence evolution accurately over the timescales needed to understand LECA [128–131]. Phylogenomic analysis is vulnerable to artefacts [13–15] and understanding the proportion of gene families for which the signal is saturated and therefore prone to artefacts will be important to consider when evaluating support for different eukaryogenesis models (**Fig 1**). As a consequence, obtaining sufficient phylogenetic resolution for many gene families adds a considerable margin of error to any estimates for the ancestry of the LECA gene repertoire. Indeed, one of the most important results stemming from any study of LECA and eukaryogenesis would be to determine what proportion of the LECA gene set is reliable for phylogenetic inference beyond the eukaryotic clade and thus potentially useful for distinguishing between alternative hypotheses of gene ancestry.

Despite more than 2 decades of research, no data sets define the gene family repertoire that would help us to reconstruct the widest characteristics of LECA and evaluate eukaryogenesis hypotheses. This limits our ability to quantitatively estimate contributions to the stem lineages between FECA(s) and LECA from different sources, through either HGT or additional endosymbiotic partners. The absence of these data also prevents us from understanding the roles of evolutionary phenomena such as de novo gene evolution, gene fusion, and gene duplication. Understanding such phenomena requires resolved data sets and detailed approaches (e.g., [132]). We therefore argue that it is not possible to rigorously address the origin of eukaryotes without a quantitative assessment of the gene repertoire of LECA.

Gene duplication—A further complexity in understanding LECA

A consideration for LECA reconstruction analyses is the accurate identification and determination of the relative contributions of gene duplication and loss [133] (i.e., paralogous gene

family expansions and differential paralog loss—see **Fig 3A**). Eukaryotes have a much greater abundance of duplicate genes and functionally differentiated paralogs than do prokaryotes, demonstrating the profound significance of this process in eukaryotic evolution, before, during, and after the divergence of the major lineages in the eukaryotic tree [48]. Indeed, paralogous expansions underpin many of the LECA cellular systems discussed in **Box 1 and Fig 2**. For example, the diversification of motor proteins through gene duplication and domain recombination has been a factor in the evolution of eukaryotic cellular complexity (e.g., [59–61,134]). Furthermore, large-scale expansions have occurred in many gene families such as small GTPases [135,136], kinases [137], and transcription factors [89,90] that control eukaryotic cellular pathways. Further back in time, gene families derived from archaea (e.g., those that play roles in DNA storage and replication and protein folding) have been subject to numerous rounds of gene duplication before LECA [76,79,138]. A full understanding of the biology of LECA thus requires an accurate delineation of the role of gene duplication before and after eukaryogenesis for both the prokaryote-derived and eukaryote-specific gene families.

Box 1: What do we know about LECA?

LECA reconstruction studies have largely focused on either cellular system-by-system analyses or investigations that take stock of total gene repertoire (e.g., [48,53]). System-specific analyses have demonstrated that LECA possessed: (i) a nucleus, nucleolus, nuclear lamina, and nuclear pore complexes [54–57]; (ii) a complex actin- and tubulin-based cytoskeleton including associated motor proteins and the systems to encode flagella, pseudo/filopodia [58–62], and mitosis encompassing a complex cell replication cycle [63–66]; (iii) genes necessary for meiosis and a facultative sexual cycle [53,67–70]; and (iv) a complex and diversified endomembrane and endomembrane trafficking system [71–74]. LECA is also inferred to have had: (v) histone/nucleosome-based chromatin with H2A, H2B, H3, and H4 paralogs and chromatin-associated catalytic functions such as methyltransferases, modification readers, and erasers [75,76], as well as SMC-based higher-level chromatin organization [77,78]; (vi) a largely archaeal-derived DNA replication system diversified by gene duplications [79,80] with some eukaryotic-specific additions (but see [32]); (vii) a spliceosome and a diversified repertoire of introns [81–86]; (viii) linear nuclear chromosomes with centromeres and telomeres [87,88] and with multi-layered regulation of gene expression [89–91]; (ix) membranes composed of fatty acid chains linked to a glycerol-3-phosphate (G3P) head group via ester bonds [92] and containing diverse sterols [93]; (x) peroxisomes [94]; and (xi) a fully integrated mitochondrial organelle similar to those found in extant lineages, with its own genome [95–100]. The population of cells that approximately constituted LECA thus had a fully fledged and elaborate eukaryotic molecular and cellular biology (**Fig 2**), not unlike many extant heterotrophic flagellated protists [17,18,101]. These patterns do not mean, however, that these core systems are immutable. Indeed, replacements, modifications, and reductions of these systems have occurred frequently across the eukaryotic tree. These include, for example, losses of flagella [102,103], peroxisomes [104], and phagocytosis [105,106], loss or radical modification of mitochondria [107–111], and the depletion of histones [112,113]. A robust LECA gene set is essential if we are to understand and appropriately account for secondary loss in eukaryotic evolution.Fig 2Cellular features inferred to be present in LECA.

This schematic follows on from [17] and summarises the cellular features discussed in the section titled "What do we know about LECA?" (and references therein). Note that

the process of meiosis, mitosis, cell division, associated machines, and processes, inferred to have been present in LECA, are not shown here. Created in BioRender. Eme, L. (2024) https://BioRender.com/w64x492. LECA, last eukaryotic common ancestor.

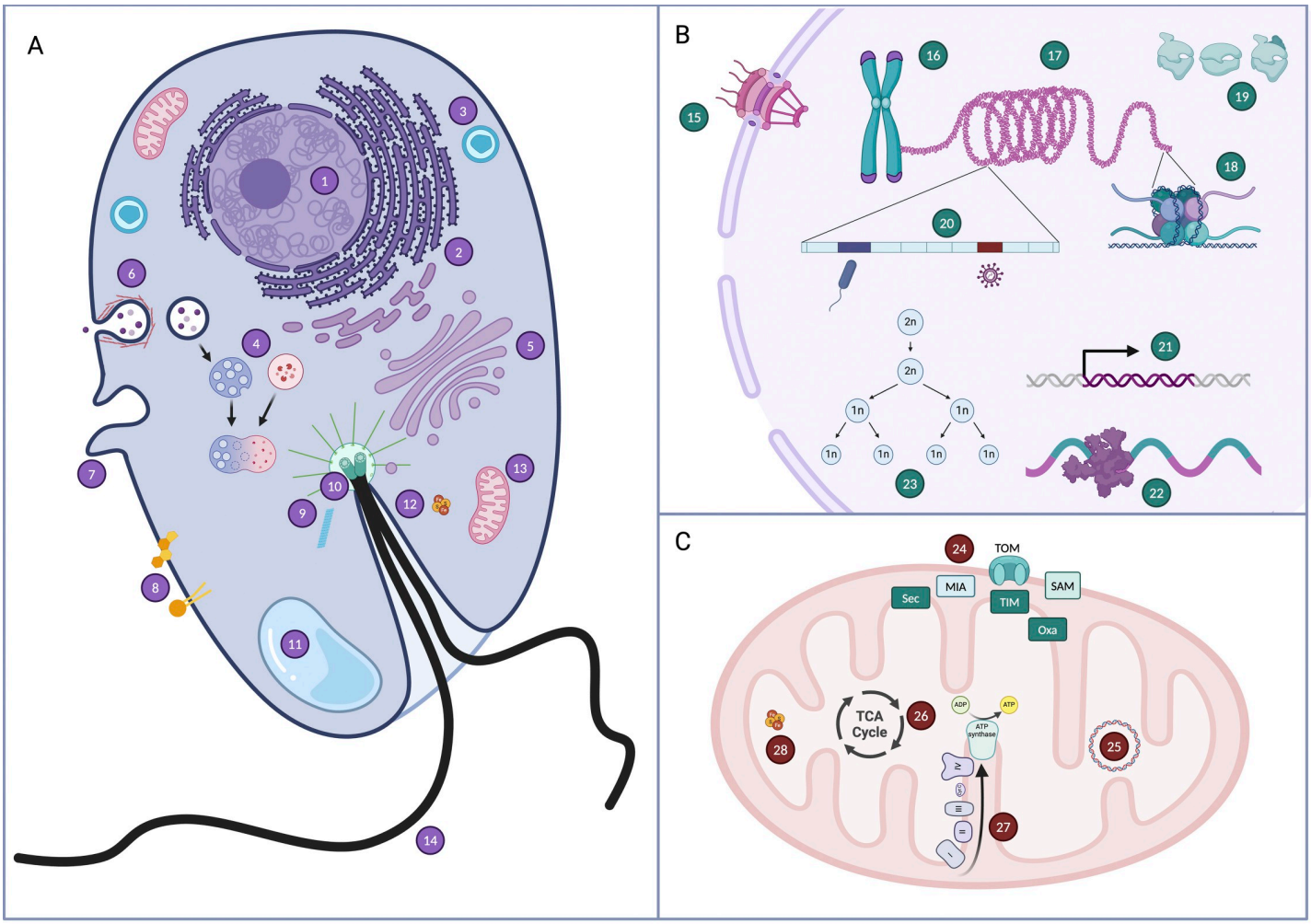

Fig 2. Cellular features inferred to be present in LECA. This schematic follows on from [17] and summarises the cellular features discussed in the section titled "What do we know about LECA?" (and references therein). Note that the process of meiosis, mitosis, cell division, associated machines, and processes, inferred to have been present in LECA, are not shown here. Created in BioRender. Eme, L. (2024) https://BioRender.com/w64x492. LECA, last eukaryotic common ancestor.

https://doi.org/10.1371/journal.pbio.3002917.g002

How to resolve LECA: A call for cooperative action, accessible data, and a path towards reconciliation of distinct data sets

A key problem in the field of eukaryotic evolution is that the inventory of genes from across the diversity of life is incomplete and requires continual updates as new lineages are discovered, more genomes are sequenced, and as annotation of existing genomes improves (e.g., [139–141]). Data sets relevant to the reconstruction of LECA will amass quickly, for example, as a product of the Earth Biogenome Project [142] and the associated Darwin Tree of Life [143] and Aquatic Symbiosis Genomics [144] projects, and from metagenomic sampling of microbial diversity (e.g., [9,145–147]). Furthermore, as models of sequence evolution continue to improve [23,128–131,148,149] phylogenetic and phylogenomic relationships will be re-evaluated. Improved homology detection methods, particularly structure-based methods utilising the latest AI approaches [150], will resolve homology relationships, trigger re-analysis of the relative contributions of different prokaryotes to LECA, and further improve comparative phylogenetic analyses [151]. Such approaches will also help to clarify patterns of homology between divergent eukaryotic genes, leading to a reassessment of when and how ortholog groups were acquired within the eukaryotic radiation. For these reasons, attempts to define a LECA gene repertoire are a "hostage to fortune"; as new data become available and methods improve; revision and tools to enable revision are needed. We provide a set of recommendations that could serve as a pathway forward and sketch an analytical approach allowing reconciliation of different LECA data sets (**Fig 3A and 3B**).

We are advocating for a large-scale, cooperative, and community-minded approach to inferring a full LECA gene set (**Box 2**). This reconstruction requires the accurate estimation of eukaryotic orthologous gene family relationships [152], followed by the identification of sister group relationships in order to identify and polarise gene duplications and, when appropriate, infer prokaryotic ancestry (**Fig 3A**). Fast approximations of ortholog clustering are possible using automated methods [153–155], but these approaches are error-prone—they can classify paralog-containing clusters as orthologs (under-splitting), separate in-paralogs/recent duplications from their bona fide orthologs (over-splitting) [156], and erroneously split orthologous groups due to high levels of sequence divergence (also over-splitting) [157]. As a consequence, some researchers combine fast ortholog clustering with manual curation [158,159], a practice that can mitigate such issues but also introduces subjectivity. The greater part of these curation process (and the subjectivity involved) is lost to the wider scientific record and can produce data sets that are difficult to analyse, compare, and critically assess [160]. Providing access to the data from these "chains" of analyses will be important, especially for the systematic integration of new data sets which allows for the revision of ortholog classifications (see **Box 2** recommendations and **Fig 3B**).

Once ortholog groups are established, it is in principle possible to compare these groups with homologous gene clusters from prokaryotes and then map the origin of eukaryotic gene families onto the prokaryotic tree of life (**Fig 3C**). Such analyses are complicated by ever-growing data sets that often result in sequence alignment sizes that restrict the use of sophisticated phylogenetic methods, in turn necessitating phylogenetically informed down-sampling. Nonetheless, ancestral state reconstruction using parsimony, Bayesian, or maximum likelihood methods [161] can be used to map gene family acquisition to a species tree, each giving somewhat different views of how gene content is inherited across the tree [162]. Some methods also allow for joint species tree/gene tree reconciliation analyses using likelihood-based inference, enabling mapping of gene repertoires onto species phylogenies [163]. Understanding the pattern of gene flow identified by these differing approaches requires further investigation of individual gene phylogenies to identify eukaryote-to-eukaryote or prokaryote-to-eukaryote HGT,

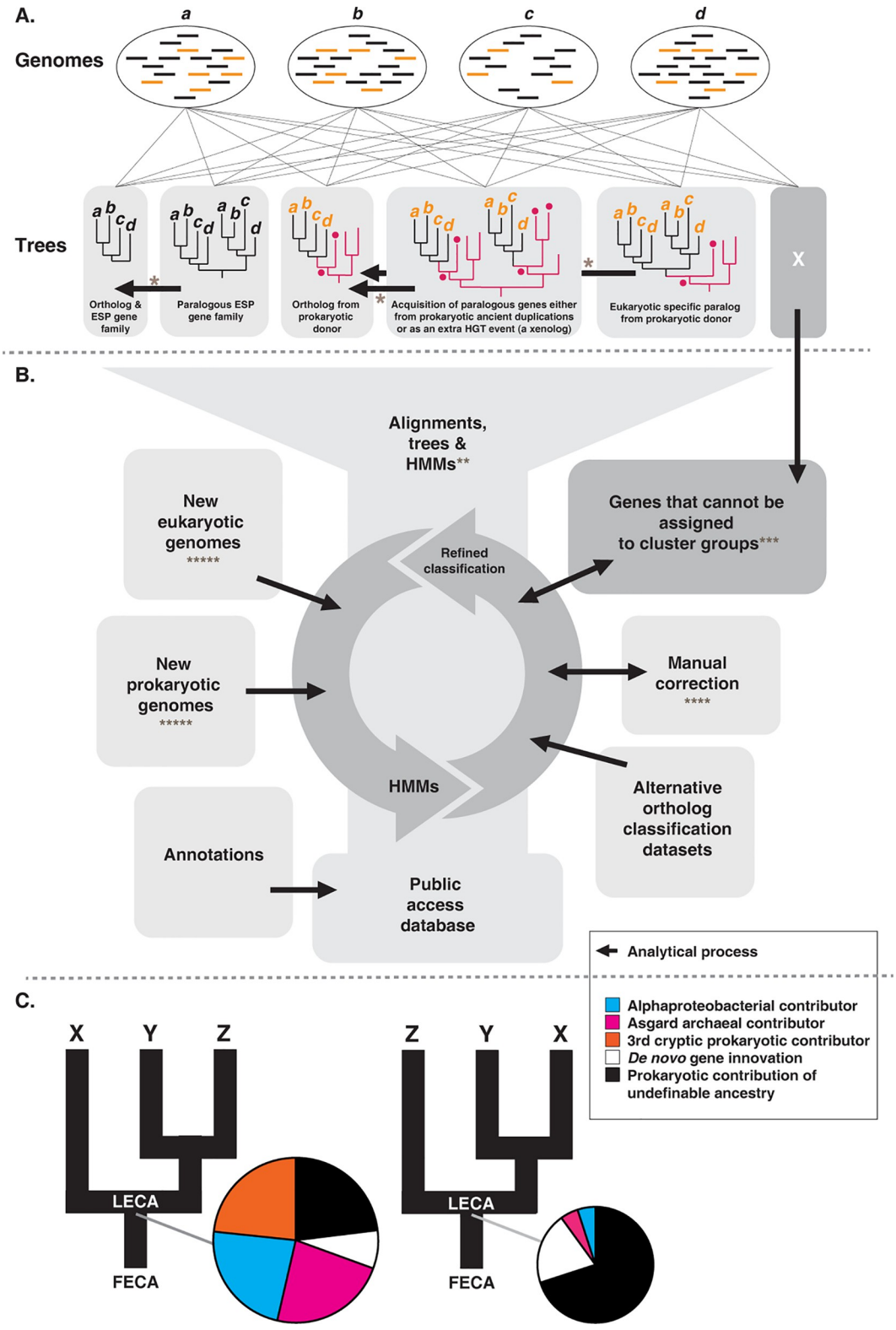

Fig 3. Proposed LECA gene repertoire analysis pipeline. (A) Eukaryotic gene complements are divided into candidate ortholog groups using phylogenetic trees. Black arrows indicate how phylogenetic analyses can be used to move from gene family phylogenies to distinct ortholog groups. Black blocks indicate genes that are specific to eukaryotes (i.e., ESPs). Orange blocks indicate eukaryotic genes of prokaryotic ancestry (phylogenetic donor-relationship is identified by red branches in the

trees; red discs on the tree indicate information for inferring provenance of prokaryotic ancestry, e.g., taxonomy and node support statistics). Note that numerous genes are likely to be classified as "genes that cannot be assigned to cluster groups" (marked as box X). This pool is a repository which would allow for further revision, addition of unclassified genes to new cluster groups as they arise, or subsequent inclusion within established cluster groups as more genome data are included and the HMMs are revised. The broader process would allow cross referencing of specific orthologs to larger gene clusters, thereby allowing the ultimate ancestry of ortholog families to be inferred. **(B)** Overview of analytical process that would allow community-based revision of ortholog cluster-groupings necessary for LECA gene repertoire estimations. This process is based on HMM generation and several levels of revision allowing cluster groupings to be updated with input from numerous additional sources of data (as shown). **(C)** LECA gene repertoire estimation based on ancestral state estimation and allowing for alternative eukaryotic species tree topologies. Sources of analytical challenge and error are marked using "*" convention. *Resolving gene clusters and ortholog groups will be a highly challenging due to lack of phylogenetic resolution and hidden paralogy, likely leading to a high proportion of genes that cannot be resolved to cluster or ortholog groups. It is for this reason we advocate for iterative chains of analysis allowing for appropriate identification of such gene sets and where possible revisions. **HMMs generated for ortholog groups will likely cross-sample paralogs and/or xenologs. New tools are needed to allow ortholog sampling that excludes paralogs (e.g., [174]). ***Pipelines to cluster orphan genes will be subject to high error with false clustering of unrelated genes. ****Manual correction will involve subjective error; this is unavoidable but community access to these processes is critical to allow for downstream improvement. *****The flow of new genomic data, with different assembly and annotation standards and varying sources of contamination, will be a difficult challenge to integrate while also maintaining standards for comparative analyses. Legend is shown in a box. ESP, eukaryote signature protein; HMM, hidden Markov model; LECA, last eukaryotic common ancestor.

https://doi.org/10.1371/journal.pbio.3002917.g003

as well as sequence data contamination, to avoid overestimating gene presence in ancestors such as LECA. The inference of sister group relationships informed by ancestral state reconstruction between prokaryotic gene clusters and eukaryotic orthologs can allow understanding of when and how gene families were acquired by eukaryotes (e.g., HGT, endosymbiosis, de novo gene acquisition, and gene duplication). Ancestral gene complements can also be reconstructed under various eukaryotic phylogenies and root hypotheses so that different ortholog gene family repertoires can be compared, which is important when there is uncertainty regarding topological relationships within the species phylogeny (e.g., [164–167]) (**Fig 3C**).

Ortholog detection is a challenge that increases in complexity with gene family size, gene loss events, and data asymmetry, e.g., the comparison of highly sampled taxonomic groups with groups with few genome sequences. The goal of the community-based "Quest for Orthologs" initiative [168] is to evaluate the strengths and weaknesses of tools for identifying orthologous gene families; members are committed to open exchange of methods and approaches supported by shared benchmarking tools enabling cross validation [169]. This is exactly the approach needed for the study of LECA, a community that provides tools and benchmarks, and sets standards for data sharing.

An aspect of this community approach would be a clear framework for systematic comparison of different LECA reconstructions. For example, given a set of putative eukaryote-wide orthologs, it is possible to generate individual hidden Markov models (HMMs [170–173]) that can be used to define each individual ortholog cluster. Current HMM methods can also be tailored to exclude certain sequences thereby allowing analyses to be targeted for specific orthologs while excluding paralogs and xenologs [174]. Refined HMM sets can then be used to compare [172] and add additional genomes to the comparative data set, allowing for iterative revision of both the ortholog groups and the HMMs themselves (**Fig 3B**).

One advantage of HMM-to-HMM comparison methods (e.g., [172]) is that they make it possible to compare and, if needed, reconcile different LECA ortholog data sets. Such an approach can be used to systematically revise ortholog gene families as new LECA data sets are released. But the data must be accessible to allow systematic comparisons (**Box 2** and **Fig 3B**). Ideally, such an endeavour would be mounted as a web-based database for the community, allowing updates and corrections. Ortholog classifications can then be improved iteratively with more data and increasing engagement (**Fig 3B**). The history of source data and the revision chain would therefore be available for each gene family so researchers could view how

ortholog assignments have progressed. In addition, the orthogroups identified as part of this community effort could be of wider use. Validated orthogroups could for example be used as an update to the KOG database that provides the core of the eukaryotic level orthology in EggNOG [175], thereby providing feedback to the larger comparative genomics community and making orthogroup classifications readily usable for gene annotation and other informatic applications.

Box 2: Aspirational standards for a community approach to LECA analyses

For both large-scale and system-specific reconstructions

- Eukaryote-wide phylogenomic analyses should make phylogenetic trees, amino acid sequence alignments, and HMMs representing gene clusters, along with the underlying methods, easily accessible (e.g., through data repository services).
- Trees, alignments, and HMMs representing gene clusters should be accessible for cross comparison, i.e., presented in tractable file formats (e.g., NEWICK, FASTA, HMMs, respectively).
- Source and assembly versions of the genome data sets used for analyses should be indicated, ideally with date of access or annotation version available.
- Sequence data decontamination processes should be described, and the resulting genome/proteome made available.
- LECA repertoire estimations should account for ancient gene duplications both in the prokaryotes (pre-LECA) and within the eukaryotes, thus separating gene families into eukaryotic ortholog clusters where possible, such that paralog relationships are identifiable.
- For each LECA gene repertoire reconstruction, eukaryotic phylogenies and root hypotheses should be clearly stated and, optimally, various alternatives should be considered so that different ancestral complements can be compared.
- If ancestral gene repertoire reconstructions are estimated, alternative approaches should be compared (e.g., Dollo parsimony, maximum likelihood [with different birth/death models], Bayesian, and reconciliation approaches).
- Different methods of ancestral gene repertoire reconstruction will provide variant estimates of eukaryote-to-eukaryote HGT. This factor should be acknowledged and targeted phylogenetic analysis to validate candidate HGT families is advised.
- Automated ortholog assessment methods should be supervised and/or validated. Correction and validation processes should be recorded in a data accessible manner (e.g., differences between processed ortholog sets should be made available).
- The process of ortholog amendments should be described.

Specific to large-scale, all-systems reconstruction of LECA

- LECA repertoire estimations should identify gene sets where there is no phylogenetic resolution or there are too few alignable sites to allow conclusive phylogenetic analyses.
- For each LECA gene repertoire reconstruction, the proportion of LECA gene families for which a prokaryotic donor can or cannot be pinpointed should be indicated.
- The approach used to account for eukaryotic paralog expansions, i.e., whether expanded eukaryotic families are counted as a single entity or individually by duplicate number in LECA should be clearly stated when assigning relative percent prokaryotic contributions to LECA.
- For hypotheses invoking multiple prokaryotic donors into LECA, the relative proportion of phylogenies which support each purported prokaryotic donor group should be indicated.
- Having established the set of prokaryotic donors to LECA, ancestral gene repertoires should be used to systematically test for the role of HGT pre-LECA.

A related consideration is that the data sets arising from LECA-scale analyses are contained within the supplementary materials of complex publications. Here, the traditional publication model fails the phylogenomic endeavour because these data are not easily accessible or standardised for systematic comparison. Such comparisons are fundamental for understanding how to improve estimation of LECA gene repertoire sets. Given the data problems outlined, as a community, we must strive for data release, accessibility and analysis standards that allow for systematic comparison.

We have provided recommendations (**Box 2**) and sketched a pathway (**Fig 3A and 3B**) to enable an accessible large-scale, all-taxa reconstruction of LECA, providing access to cross-comparison and facilitating iterative improvement. To enable this endeavour, we also advocate for the development of web-based database resources to support such interactions (e.g., [176]). As genome sampling increases and ortholog sets are corrected, LECA gene complement estimation could be iteratively revised (e.g., LECA 2.0, etc.; ideally with a release schedule outlined so researchers in the field can plan accordingly). Now is the time for the community to start building LECA-specific tools and resources for handing the complicated task of data analysis required to resolve the gene repertoire of LECA in a way that caters to differential approaches and perspectives while also making iterative chains of phylogenomic analyses available. We recognise that many groups will continue with focused analyses of individual cellular systems. These analyses will complement, and can integrate with, any large-scale LECA reconstruction, providing important ground-truthing data sets for the annotation and manual correction phases outlined in **Fig 3B**. Furthermore, large-scale LECA reconstruction will identify groups of genes that are especially difficult to resolve using bioinformatic pathway-based approaches which therefore need focused analyses, making these approaches both complementary and iterative. Many of the recommendations regarding data sharing, standards, and transparency (**Box 2**) apply equally to both types of effort.

Beyond eukaryogenesis—The wider value of reconstructing LECA

The origin of the eukaryotic cell laid the foundation for a vast diversification of biological forms leading to additional major evolutionary transitions. As a resolved LECA gene repertoire provides a baseline from which to infer lineage-specific evolutionary changes within the eukaryotes, this data set will allow researchers to address a multitude of questions, both evolutionary and cell biological in nature.

A LECA gene set will allow study of the evolutionary dynamics during the early diversification of the major eukaryotic groups, including the contributions of gene gain, loss, duplication, HGT, and domain rearrangement (i.e., gene-fusions and -fissions). Such data will also support a range of downstream analyses, for example, providing expected ortholog distribution maps for evaluating eukaryotic genome assembly completion, similar to the approaches applied in BUSCO [177] and OMArk [178]. Resolved ortholog relationships will also be an important resource for concatenated multi-gene phylogenomic analysis (e.g., [179]) underpinning further investigations of the eukaryotic tree. Finally, a LECA repertoire provides a starting gene repertoire from which to infer the evolution of nearly all extant eukaryotic cellular functions. This includes the origin and spread of photosynthetic organelles [180,181], the repeated evolution of pathogenicity (e.g., [182–185]), and the multiple origins of multicellular forms such as plants, animals, fungi, and seaweeds [186].

The LECA gene set should also serve as baseline data for fundamental cell biological inquiries aiming to move beyond standard model organisms (e.g., yeast, animal, or plant). Such organisms are unrepresentative of the diversity of eukaryotic traits and cellular forms, although comparison of the 3 groups is, of course, important. The genes, proteins, and processes found in LECA can be considered ancient and are potentially generalizable as features of "the eukaryotic cell." Furthermore, the LECA analyses proposed here would identify conserved gene families present across the eukaryotes for which there is no known functional annotation. Many of these may turn out to be jötnarlogs—genes with patchy distributions, absent in model organisms, but present in diverse organisms of medical or ecological importance (e.g., [71,187]). Such data are important, for example, when researchers wish to identify a gene present in a group of pathogens/parasites with no host-encoded homologous protein as a putative drug target. Finally, a LECA gene repertoire facilitates investigation of co-occurrence patterns between uncharacterised core systems and known cell functions (e.g., [102,188]), thereby providing clues regarding function. The results of a wide range of LECA analyses can be compared to large-scale knockout libraries in model systems providing further information on function and evolution [189].

Conclusion

Resolving the early evolution of the eukaryotic cell remains a huge challenge [21]. Given its importance and antiquity, we have more hypotheses than definitive data. Consequently, every detail upon which a consensus is reached can push inferences towards one eukaryogenesis scenario over another, or help us to resolve a key factor in the early evolution of eukaryotes. An estimation of the LECA gene repertoire is a foundational data set for testing pivotal ideas about how the early eukaryotic cell evolved, providing an end state at which all eukaryogenesis models need to arrive and a starting point for understanding the evolution of major eukaryotic groups and their cellular systems. A community-wide effort to define LECA in terms of cell biology and gene repertoire will permit informed comparisons of different models so that they can be judged on their relative merits. This is a complex task, one in which different approaches and new data can radically alter patterns. Such investigation can therefore only realistically move forward through systematic community engagement with adherence to

shared standards. To that end, we have outlined recommendations for data analyses and accessibility to allow for systematic comparisons. We have also sketched out an analytical pathway that would allow for the cross comparison of LECA data sets given the changing availability of data (Fig 3). Our hope is that this framework will be useful for individual research teams and discipline-wide consortia alike, and that the ideas presented herein about how these data should and could be used will trigger new ways of thinking about the problem of eukaryogenesis and early eukaryotic cell evolution (Box 2).

Acknowledgments

The planning and writing of this article were greatly facilitated by the Moore-Simons Project on the Origin of the Eukaryotic Cell funded by the Gordon and Betty Moore Foundation and the Simons Foundation.

Author Contributions

Conceptualization: Thomas A. Richards, Joel B. Dacks.

Visualization: Thomas A. Richards, Laura Eme.

Writing – original draft: Thomas A. Richards, Laura Eme, John M. Archibald, Andrew J. Roger, Joel B. Dacks, Jeremy G. Wideman.

Writing – review & editing: Thomas A. Richards, Laura Eme, John M. Archibald, Guy Leonard, Susana M. Coelho, Alex de Mendoza, Christophe Dessimoz, Pavel Dolezal, Lillian K. Fritz-Laylin, Toni Gabaldón, Vladimír Hampl, Geert J. P. L. Kops, Michelle M. Leger, Purificacion Lopez-Garcia, James O. McInerney, David Moreira, Sergio A. Muñoz-Gómez, Daniel J. Richter, Iñaki Ruiz-Trillo, Alyson E. Santoro, Arnau Sebé-Pedrós, Berend Snel, Courtney W. Stairs, Eelco C. Tromer, Jolien J. E. van Hooff, Bill Wickstead, Tom A. Williams, Andrew J. Roger, Joel B. Dacks, Jeremy G. Wideman.

References

1. Stanier RY, Doudoroff M, Adelberg EA. The microbial world: Prentice-Hall; 1957.
2. Lane N, Martin W. The energetics of genome complexity. Nature. 2010; 467(7318):929–934. https://doi.org/10.1038/nature09486 PMID: 20962839.
3. Booth A, Doolittle WF. Eukaryogenesis, how special really? Proc Natl Acad Sci U S A. 2015; 112 (33):10278–10285. https://doi.org/10.1073/pnas.1421376112 PMID: 25883267
4. Booth A, Doolittle WF. Reply to Lane and Martin: Being and becoming eukaryotes. Proc Natl Acad Sci U S A. 2015;112(35):E4824–E. https://doi.org/10.1073/pnas.1513285112 PMID: 26283404
5. Lane N, Martin WF. Eukaryotes really are special, and mitochondria are why. Proc Natl Acad Sci U S A. 2015; 112(35):E4823–E. https://doi.org/10.1073/pnas.1509237112 PMID: 26283405
6. López-García P, Moreira D. Open questions on the origin of eukaryotes. Trends Ecol Evol. 2015; 30 (11):697–708. https://doi.org/10.1016/j.tree.2015.09.005 PMID: 26455774
7. Purificación L-G, David M. The symbiotic origin of the eukaryotic cell. C R Biol. 2023; 346:55–73. https://doi.org/10.5802/crbiol.118 PMID: 37254790
8. Eme L, Tamarit D, Caceres EF, Stairs CW, De Anda V, Schön ME, 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 PMID: 37316666
9. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015; 521(7551):173–9. Epub 20150506. https://doi.org/10.1038/nature14447 PMID: 25945739; PubMed Central PMCID: PMC4444528.
10. 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). Epub 20140801. https://doi.org/10.1101/cshperspect.a016139 PMID: 25085908; PubMed Central PMCID: PMC4107988.

11. Strassert JFH, Irisarri I, Williams TA, Burki F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat Commun. 2021; 12(1):1879. https://doi.org/10.1038/s41467-021-22044-z PMID: 33767194

12. 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):1556–62. Epub 20180820. https://doi.org/10.1038/s41559-018-0644-x PMID: 30127539; PubMed Central PMCID: PMC6152910.

13. Philippe H, Vienne DM, Ranwez V, Roure B, Baurain D, Delsuc F. Pitfalls in supermatrix phylogenomics. Eur J Taxon. 2017;0(283). https://doi.org/10.5852/ejt.2017.283

14. Simion P, Delsuc F, Philippe H. To what extent current limits of phylogenomics can be overcome? In: Scornavacca C, Delsuc F, Galtier N, editors. Phylogenetics in the Genomic Era: No commercial publisher | Authors open access book; 2020. p. 2.1:—2.1:34.

15. Steenwyk JL, Li Y, Zhou X, Shen XX, Rokas A. Incongruence in the phylogenomics era. Nat Rev Genet. 2023; 24(12):834–50. Epub 20230627. https://doi.org/10.1038/s41576-023-00620-x PMID: 37369847.

16. Dacks JB, Field MC, Buick R, Eme L, Gribaldo S, Roger AJ, et al. The changing view of eukaryogenesis–fossils, cells, lineages and how they all come together. J Cell Sci. 2016; 129(20):3695–3703. https://doi.org/10.1242/jcs.178566 PMID: 27672020

17. Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB, Field MC. Molecular paleontology and complexity in the last eukaryotic common ancestor. Crit Rev Biochem Mol Biol. 2013; 48 (4):373–396. https://doi.org/10.3109/10409238.2013.821444 PMID: 23895660; PubMed Central PMCID: PMC3791482.

18. O'Malley MA, Leger MM, Wideman JG, Ruiz-Trillo I. Concepts of the last eukaryotic common ancestor. Nat Ecol Evol. 2019; 3(3):338–344. https://doi.org/10.1038/s41559-019-0796-3 PMID: 30778187

19. Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. Nature. 2013; 504(7479):231–236. https://doi.org/10.1038/nature12779 PMID: 24336283

20. Cox CJ, Foster PG, Hirt RP, Harris SR, Embley TM. The archaebacterial origin of eukaryotes. Proc Natl Acad Sci U S A. 2008; 105(51):20356–20361. https://doi.org/10.1073/pnas.0810647105 PMID: 19073919

21. Vosseberg J, van Hooff JJE, Köstlbacher S, Panagiotou K, Tamarit D, Ettema TJG. The emerging view on the origin and early evolution of eukaryotic cells. Nature. 2024; 633:295–305. https://doi.org/10.1038/s41586-024-07677-6 PMID: 39261613

22. Bonen L, Cunningham RS, Gray MW, Doolittle WF. Wheat embryo mitochondrial 18S ribosomal RNA: evidence for its prokaryotic nature. Nucleic Acids Res. 1977; 4(3):663–671. https://doi.org/10.1093/nar/4.3.663 866186; PubMed Central PMCID: PMC342470. PMID: 866186

23. Muñoz-Gómez SA, Susko E, Williamson K, Eme L, Slamovits CH, Moreira D, et al. 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 PMID: 35027725

24. Martijn J, Vosseberg J, Guy L, Offre P, Ettema TJG. Deep mitochondrial origin outside the sampled alphaproteobacteria. Nature. 2018; 557(7703):101–5. Epub 20180425. https://doi.org/10.1038/s41586-018-0059-5 PMID: 29695865.

25. Gabaldón T. Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis. IUBMB Life. 2018; 70(12):1188–1196. https://doi.org/10.1002/iub.1950 PMID: 30358047

26. López-García P, Moreira D. The syntrophy hypothesis for the origin of eukaryotes revisited. Nat Microbiol. 2020; 5(5):655–67. Epub 20200427. https://doi.org/10.1038/s41564-020-0710-4 PMID: 32341569.

27. Stairs CW, Dharamshi JE, Tamarit D, Eme L, Jørgensen SL, Spang A, et al. Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. Sci Adv. 2020; 6(35):eabb7258. https://doi.org/10.1126/sciadv.abb7258 PMID: 32923644

28. Devos DP, Reynaud EG. Intermediate steps. Science. 2010; 330(6008):1187–1188. https://doi.org/10.1126/science.1196720 PMID: 21109658.

29. Bell PJL. Eukaryogenesis: the rise of an emergent superorganism. Front Microbiol. 2022: 13. https://doi.org/10.3389/fmicb.2022.858064 PMID: 35633668

30. Moreira D, López-García P. Evolution of viruses and cells: do we need a fourth domain of life to explain the origin of eukaryotes? Philos Trans R Soc Lond B Biol Sci. 2015; 370(1678):20140327. https://doi.org/10.1098/rstb.2014.0327 PMID: 26323758

31. Forterre P, Gaïa M. Giant viruses and the origin of modern eukaryotes. Curr Opin Microbiol. 2016; 31:44–49. https://doi.org/10.1016/j.mib.2016.02.001 PMID: 26894379

32. Karki S, Barth ZK, Aylward FO. Chimeric origin of eukaryotes from Asgard archaea and ancestral giant viruses. bioRxiv. 2024:2024.04.22.590592. https://doi.org/10.1101/2024.04.22.590592

33. Bell PJL. Evidence supporting a viral origin of the eukaryotic nucleus. Virus Res. 2020; 289:198168. https://doi.org/10.1016/j.virusres.2020.198168 PMID: 32961211

34. Irwin NAT, Pittis AA, Richards TA, Keeling PJ. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nat Microbiol. 2022; 7(2):327–336. https://doi.org/10.1038/s41564-021-01026-3 PMID: 34972821

35. Bonen L, Doolittle WF. On the prokaryotic nature of red algal chloroplasts. Proc Natl Acad Sci U S A. 1975; 72(6):2310–2314. https://doi.org/10.1073/pnas.72.6.2310 PMID: 1056032

36. Martin W, Müller M. The hydrogen hypothesis for the first eukaryote. Nature. 1998; 392(6671):37–41. https://doi.org/10.1038/32096 PMID: 9510246

37. Baum DA, Baum B. An inside-out origin for the eukaryotic cell. BMC Biol. 2014; 12(1):76. https://doi.org/10.1186/s12915-014-0076-2 PMID: 25350791

38. Imachi H, Nobu MK, Nakahara N, Morono Y, Ogawara M, Takaki Y, et al. Isolation of an archaeon at the prokaryote-eukaryote interface. Nature. 2020; 577(7791):519–25. Epub 20200115. https://doi.org/10.1038/s41586-019-1916-6 PMID: 31942073; PubMed Central PMCID: PMC7015854.

39. Martijn J, Ettema TJ. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. Biochem Soc Trans. 2013; 41(1):451–457. https://doi.org/10.1042/BST20120292 PMID: 23356327.

40. Margulis L. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. Proc Natl Acad Sci U S A. 1996; 93(3):1071–1076. https://doi.org/10.1073/pnas.93.3.1071 PMID: 8577716

41. Cavalier-Smith T. Ciliary transition zone evolution and the root of the eukaryote tree: implications for opisthokont origin and classification of kingdoms Protozoa, Plantae, and Fungi. Protoplasma. 2022; 259(3):487–593. Epub 20211223. https://doi.org/10.1007/s00709-021-01665-7 PMID: 34940909; PubMed Central PMCID: PMC9010356.

42. Martin W, Hoffmeister M, Rotte C, Henze K. An overview of endosymbiotic models for the origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and their heterotrophic lifestyle. Biol Chem. 2001; 382(11):1521–1539. https://doi.org/10.1515/BC.2001.187 PMID: 11767942.

43. McInerney JO, Martin WF, Koonin EV, Allen JF, Galperin MY, Lane N, et al. Planctomycetes and eukaryotes: a case of analogy not homology. Bioessays. 2011; 33(11):810–7. Epub 20110822. https://doi.org/10.1002/bies.201100045 PMID: 21858844; PubMed Central PMCID: PMC3795523.

44. Donoghue PCJ, Kay C, Spang A, Szöllősi G, Nenarokova A, Moody ERR, et al. Defining eukaryotes to dissect eukaryogenesis. Curr Biol. 2023; 33(17):R919–R929. https://doi.org/10.1016/j.cub.2023.07.048 PMID: 37699353

45. Raval PK, Garg SG, Gould SB. Endosymbiotic selective pressure at the origin of eukaryotic cell biology. Elife. 2022; 11:e81033. https://doi.org/10.7554/eLife.81033 PMID: 36355038

46. Eme L, Spang A, Lombard J, Stairs CW, Ettema TJG. Archaea and the origin of eukaryotes. Nat Rev Microbiol. 2017; 15(12):711–723. https://doi.org/10.1038/nrmicro.2017.133 PMID: 29123225.

47. 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 PMID: 26840490

48. Vosseberg J, van Hooff JJE, Marcet-Houben M, van Vlimmeren A, van Wijk LM, Gabaldón T, et al. Timing the origin of eukaryotic cellular complexity with ancient duplications. Nat Ecol Evol. 2021; 5(1):92–100. https://doi.org/10.1038/s41559-020-01320-z PMID: 33106602

49. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Bäckström D, Juzokaite L, Vancaester E, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017; 541(7637):353–358. https://doi.org/10.1038/nature21031 PMID: 28077874

50. Lazcano A, Peretó J. Prokaryotic symbiotic consortia and the origin of nucleated cells: a critical review of Lynn Margulis hypothesis. Biosystems. 2021; 204:104408. https://doi.org/10.1016/j.biosystems.2021.104408 PMID: 33744400

51. Rochette NC, Brochier-Armanet C, Gouy M. Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. Mol Biol Evol. 2014; 31(4):832–45. Epub 20140107. https://doi.org/10.1093/molbev/mst272 PMID: 24398320; PubMed Central PMCID: PMC3969559.

52. Barrera-Redondo J, Lotharukpong JS, Drost HG, Coelho SM. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. Genome Biol. 2023; 24(1):54. Epub 20230324. https://doi.org/10.1186/s13059-023-02895-z PMID: 36964572; PubMed Central PMCID: PMC10037820.

53. Newman D, Whelan FJ, Moore M, Rusilowicz M, McInerney JO. Reconstructing and analysing the genome of the last eukaryote common ancestor to better understand the transition from FECA to LECA. bioRxiv. 2019:538264. https://doi.org/10.1101/538264

54. Koreny L, Field MC. Ancient eukaryotic origin and evolutionary plasticity of nuclear lamina. Genome Biol Evol. 2016; 8(9):2663–2671. https://doi.org/10.1093/gbe/evw087 PMID: 27189989

55. Mans B, Anantharaman V, Aravind L, Koonin EV. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore pomplex. Cell Cycle. 2004; 3(12):1625–1650. https://doi.org/10.4161/cc.3.12.1316 PMID: 15611647

56. Makarov AA, Padilla-Mejia NE, Field MC. Evolution and diversification of the nuclear pore complex. Biochem Soc Trans. 2021; 49(4):1601–1619. https://doi.org/10.1042/BST20200570 PMID: 34282823; PubMed Central PMCID: PMC8421043.

57. Neumann N, Lundin D, Poole AM. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. PLoS ONE. 2010; 5(10):e13241. https://doi.org/10.1371/journal.pone.0013241 PMID: 20949036

58. Wickstead B, Gull K. The evolution of the cytoskeleton. J Cell Biol. 2011; 194(4):513–525. https://doi.org/10.1083/jcb.201102065 PMID: 21859859; PubMed Central PMCID: PMC3160578.

59. Richards TA, Cavalier-Smith T. Myosin domain evolution and the primary divergence of eukaryotes. Nature. 2005; 436(7054):1113–1118. https://doi.org/10.1038/nature03949 PMID: 16121172

60. Wickstead B, Gull K. Dyneins across eukaryotes: a comparative genomic analysis. Traffic. 2007; 8 (12):1708–21. Epub 20070926. https://doi.org/10.1111/j.1600-0854.2007.00646.x PMID: 17897317; PubMed Central PMCID: PMC2239267.

61. Wickstead B, Gull K, Richards TA. Patterns of kinesin evolution reveal a complex ancestral eukaryote with a multifunctional cytoskeleton. BMC Evol Biol. 2010; 10:110. Epub 20100427. https://doi.org/10.1186/1471-2148-10-110 PMID: 20423470; PubMed Central PMCID: PMC2867816.

62. Velle KB, Fritz-Laylin LK. Diversity and evolution of actin-dependent phenotypes. Curr Opin Genet Dev. 2019;58–59:40–8. Epub 20190826. https://doi.org/10.1016/j.gde.2019.07.016 PMID: 31466039.

63. Eme L, Moreira D, Talla E, Brochier-Armanet C. A complex cell division machinery was present in the last common ancestor of eukaryotes. PLoS ONE. 2009; 4(4):e5021. https://doi.org/10.1371/journal.pone.0005021 PMID: 19352429

64. Eme L, Trilles A, Moreira D, Brochier-Armanet C. The phylogenomic analysis of the anaphase promoting complex and its targets points to complex and modern-like control of the cell cycle in the last common ancestor of eukaryotes. BMC Evol Biol. 2011; 11(1):265. https://doi.org/10.1186/1471-2148-11-265 PMID: 21943402

65. Tromer EC, van Hooff JJE, Kops G, Snel B. Mosaic origin of the eukaryotic kinetochore. Proc Natl Acad Sci U S A. 2019; 116(26):12873–82. Epub 20190524. https://doi.org/10.1073/pnas.1821945116 PMID: 31127038; PubMed Central PMCID: PMC6601020.

66. van Hooff JJ, Tromer E, van Wijk LM, Snel B, Kops GJ. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. EMBO Rep. 2017; 18(9):1559–71. Epub 20170622. https://doi.org/10.15252/embr.201744102 PMID: 28642229; PubMed Central PMCID: PMC5579357.

67. Malik SB, Ramesh MA, Hulstrand AM, Logsdon JM Jr. Protist homologs of the meiotic Spo11 gene and topoisomerase VI reveal an evolutionary history of gene duplication and lineage-specific loss. Mol Biol Evol. 2007; 24(12):2827–41. Epub 20071005. https://doi.org/10.1093/molbev/msm217 PMID: 17921483.

68. Ramesh MA, Malik SB, Logsdon JM Jr. A phylogenomic inventory of meiotic genes; evidence for sex in *Giardia* and an early eukaryotic origin of meiosis. Curr Biol. 2005; 15(2):185–191. https://doi.org/10.1016/j.cub.2005.01.003 PMID: 15668177.

69. Wilkins AS, Holliday R. The evolution of meiosis from mitosis. Genetics. 2009; 181(1):3–12. https://doi.org/10.1534/genetics.108.099762 PMID: 19139151; PubMed Central PMCID: PMC2621177.

70. Hurst LD, Nurse P. A note on the evolution of meiosis. J Theor Biol. 1991; 150(4):561–563. https://doi.org/10.1016/s0022-5193(05)80447-3 PMID: 1943134

71. More K, Klinger CM, Barlow LD, Dacks JB. Evolution and natural history of membrane trafficking in eukaryotes. Biol Rev. 2020; 30(10):R553–R564. https://doi.org/10.1016/j.cub.2020.03.068 PMID: 32428497

72. Prokopchuk G, Butenko A, Dacks JB, Speijer D, Field MC, Lukeš J. Lessons from the deep: mechanisms behind diversification of eukaryotic protein complexes. Biol Rev. 2023; 98(6):1910–1927. https://doi.org/10.1111/brv.12988 PMID: 37336550

73. Jansen RLM, Santana-Molina C, van den Noort M, Devos DP, van der Klei IJ. Comparative genomics of peroxisome biogenesis proteins: making sense of the PEX proteins. Front Cell Dev Biol. 2021;

9:654163. Epub 20210520. https://doi.org/10.3389/fcell.2021.654163 PMID: 34095119; PubMed Central PMCID: PMC8172628.

74. Gabaldón T. Evolution of the peroxisomal proteome. Subcell Biochem. 2018; 89:221–233. https://doi.org/10.1007/978-981-13-2233-4_9 PMID: 30378025.

75. Grau-Bové X, Navarrete C, Chiva C, Pribasnig T, Antó M, Torruella G, et al. A phylogenetic and proteomic reconstruction of eukaryotic chromatin evolution. Nat Ecol Evol. 2022; 6(7):1007–1023. https://doi.org/10.1038/s41559-022-01771-6 PMID: 35680998

76. Irwin NA, Richards TA. Self-assembling viral histones are evolutionary intermediates between archaeal and eukaryotic nucleosomes. Nat Microbiol. 2024:1–12.

77. van Hooff JJE, Raas MWD, Tromer EC, Eme L. Shaping up genomes: prokaryotic roots and eukaryotic diversification of SMC complexes. bioRxiv. 2024:2024.01.07.573240. https://doi.org/10.1101/2024.01.07.573240

78. Yoshinaga M, Inagaki Y. Ubiquity and origins of structural maintenance of chromosomes (SMC) proteins in eukaryotes. Genome Biol Evol. 2021; 13(12):evab256. https://doi.org/10.1093/gbe/evab256 PMID: 34894224

79. Aves SJ, Liu Y, Richards TA. Evolutionary diversification of eukaryotic DNA replication machinery. Subcell Biochem. 2012; 62:19–35. https://doi.org/10.1007/978-94-007-4572-8_2 PMID: 22918578.

80. Liu Y, Richards TA, Aves SJ. Ancient diversification of eukaryotic MCM DNA replication proteins. BMC Evol Biol. 2009; 9:60. Epub 20090317. https://doi.org/10.1186/1471-2148-9-60 PMID: 19292915; PubMed Central PMCID: PMC2667178.

81. Vosseberg J, Schinkel M, Gremmen S, Snel B. The spread of the first introns in proto-eukaryotic paralogs. Comun Biol. 2022; 5(1):476. https://doi.org/10.1038/s42003-022-03426-5 PMID: 35589959

82. Vosseberg J, Stolker D, von der Dunk SHA, Snel B. Integrating phylogenetics with intron positions illuminates the origin of the complex spliceosome. Mol Biol Evol. 2023; 40(1):msad011. https://doi.org/10.1093/molbev/msad011 PMID: 36631250

83. Koonin EV. Intron-dominated genomes of early ancestors of eukaryotes. J Hered. 2009; 100(5):618–623. https://doi.org/10.1093/jhered/esp056 PMID: 19617525

84. Simpson AGB, MacQuarrie EK, Roger AJ. Early origin of canonical introns. Nature. 2002; 419 (6904):270. https://doi.org/10.1038/419270a PMID: 12239559

85. Collins L, Penny D. Complex spliceosomal organization ancestral to extant eukaryotes. Mol Biol Evol. 2005; 22(4):1053–66. Epub 20050119. https://doi.org/10.1093/molbev/msi091 PMID: 15659557.

86. Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct. 2006; 1(1):22. https://doi.org/10.1186/1745-6150-1-22 PMID: 16907971

87. Schrumpfová PP, Fajkus J. Composition and function of telomerase—a polymerase associated with the origin of eukaryotes. Biomol [Internet]. 2020; 10(10). https://doi.org/10.3390/biom10101425 PMID: 33050064

88. de Lange T. A loopy view of telomere evolution. Front Genet. 2015:6. https://doi.org/10.3389/fgene.2015.00321 PMID: 26539211

89. de Mendoza A, Sebé-Pedrós A. Origin and evolution of eukaryotic transcription factors. Curr Opin Genet Dev. 2019;58–59:25–32. Epub 20190826. https://doi.org/10.1016/j.gde.2019.07.010 PMID: 31466037.

90. de Mendoza A, Sebé-Pedrós A, Šestak MS, Matejcic M, Torruella G, Domazet-Loso T, et al. Transcription factor evolution in eukaryotes and the assembly of the regulatory toolkit in multicellular lineages. Proc Natl Acad Sci U S A. 2013; 110(50):E4858–66. Epub 20131125. https://doi.org/10.1073/pnas.1311818110 PMID: 24277850; PubMed Central PMCID: PMC3864300.

91. Iyer LM, Anantharaman V, Wolf MY, Aravind L. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. Int J Parasitol. 2008; 38(1):1–31. Epub 20070915. https://doi.org/10.1016/j.ijpara.2007.07.018 PMID: 17949725.

92. Lombard J, López-García P, Moreira D. The early evolution of lipid membranes and the three domains of life. Nat Rev Microbiol. 2012; 10(7):507–515. https://doi.org/10.1038/nrmicro2815 PMID: 22683881

93. Desmond E, Gribaldo S. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. Genome Biol Evol. 2009; 1:364–81. Epub 20090910. https://doi.org/10.1093/gbe/evp036 PMID: 20333205; PubMed Central PMCID: PMC2817430.

94. Gabaldón T. Peroxisome diversity and evolution. Philos Trans R Soc Lond B Biol Sci. 2010; 365 (1541):765–773. https://doi.org/10.1098/rstb.2009.0240 PMID: 20124343; PubMed Central PMCID: PMC2817229.

95. Roger AJ, Muñoz-Gómez SA, Kamikawa R. The origin and diversification of mitochondria. Curr Biol. 2017; 27(21):R1177–R1192. https://doi.org/10.1016/j.cub.2017.09.015 PMID: 29112874

96. Gabaldón T, Huynen MA. From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. PLoS Comput Biol. 2007; 3(11):e219. https://doi.org/10.1371/journal.pcbi.0030219 PMID: 17983265

97. Sinha SD, Wideman JG. The persistent homology of mitochondrial ATP synthases. iScience. 2023; 26(5):106700. https://doi.org/10.1016/j.isci.2023.106700 PMID: 37250340

98. Mani J, Meisinger C, Schneider A. Peeping at TOMs—diverse entry gates to mitochondria provide insights into the evolution of eukaryotes. Mol Biol Evol. 2016; 33(2):337–351. https://doi.org/10.1093/molbev/msv219 PMID: 26474847

99. Butenko A, Lukeš J, Speijer D, Wideman JG. Mitochondrial genomes revisited: why do different lineages retain different genes? BMC Biol. 2024; 22(1):15. https://doi.org/10.1186/s12915-024-01824-1 PMID: 38273274

100. Petrů M, Dohnálek V, Füssy Z, Doležal P. Fates of Sec, Tat, and YidC translocases in mitochondria and other eukaryotic compartments. Mol Biol Evol. 2021; 38(12):5241–5254. https://doi.org/10.1093/molbev/msab253 PMID: 34436602; PubMed Central PMCID: PMC8662606.

101. Torruella G, Galindo LJ, Moreira D, López-García P. Phylogenomics of neglected flagellated protists supports a revised eukaryotic tree of life. bioRxiv. 2024:2024.05.15.594285. https://doi.org/10.1101/2024.05.15.594285

102. Wickstead B, Gull K. A "holistic" kinesin phylogeny reveals new kinesin families and predicts protein functions. Mol Biol Cell. 2006; 17(4):1734–43. Epub 20060215. https://doi.org/10.1091/mbc.e05-11-1090 PMID: 16481395; PubMed Central PMCID: PMC1415282.

103. Moran J, McKean PG, Ginger ML. Eukaryotic flagella: variations in form, function, and composition during evolution. Bioscience. 2014; 64(12):1103–1114. https://doi.org/10.1093/biosci/biu175

104. Žárský V, Tachezy J. Evolutionary loss of peroxisomes–not limited to parasites. Biol Direct. 2015; 10 (1):74. https://doi.org/10.1186/s13062-015-0101-6 PMID: 26700421

105. Merényi Z, Krizsán K, Sahu N, Liu X-B, Bálint B, Stajich JE, et al. Genomes of fungi and relatives reveal delayed loss of ancestral gene families and evolution of key fungal traits. Nat Ecol Evol. 2023; 7 (8):1221–1231. https://doi.org/10.1038/s41559-023-02095-9 PMID: 37349567

106. Richards TA, Leonard G, Wideman JG. What defines the "kingdom" Fungi? Microbiol Spectr. 2017; 5 (3):10.1128/microbiolspec.funk-0044-2017. https://doi.org/10.1128/microbiolspec.FUNK-0044-2017 PMID: 28643626

107. Moreira D, Blaz J, Kim E, Eme L. A gene-rich mitochondrion with a unique ancestral protein transport system. bioRxiv. 2024:2024.01.30.577968. https://doi.org/10.1016/j.cub.2024.07.017 PMID: 39084221

108. Novák LVF, Treitli SC, Pyrih J, Hałakuc P, Pipaliya SV, Vacek V, et al. Genomics of preaxostyla flagellates illuminates the path towards the loss of mitochondria. PLoS Genet. 2023; 19(12):e1011050. https://doi.org/10.1371/journal.pgen.1011050 PMID: 38060519

109. Karnkowska A, Vacek V, Zubáčová Z, Treitli SC, Petrželková R, Eme L, et al. A eukaryote without a mitochondrial organelle. Curr Biol. 2016; 26(10):1274–84. Epub 20160512. https://doi.org/10.1016/j.cub.2016.03.053 PMID: 27185558.

110. Fukasawa Y, Oda T, Tomii K, Imai K. Origin and evolutionary alteration of the mitochondrial import system in eukaryotic lineages. Mol Biol Evol. 2017; 34(7):1574–1586. https://doi.org/10.1093/molbev/msx096 PMID: 28369657; PubMed Central PMCID: PMC5455965.

111. Burger G, Gray MW, Forget L, Lang BF. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout Jakobid protists. Genome Biol Evol. 2013; 5(2):418–438. https://doi.org/10.1093/gbe/evt008 PMID: 23335123

112. Irwin NAT, Martin BJE, Young BP, Browne MJG, Flaus A, Loewen CJR, et al. Viral proteins as a potential driver of histone depletion in dinoflagellates. Nat Commun. 2018; 9(1):1535. Epub 20180418. https://doi.org/10.1038/s41467-018-03993-4 PMID: 29670105; PubMed Central PMCID: PMC5906630.

113. Gornik SG, Ford KL, Mulhern TD, Bacic A, McFadden GI, Waller RF. Loss of nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in dinoflagellates. Curr Biol. 2012; 22 (24):2303–2312. https://doi.org/10.1016/j.cub.2012.10.036 PMID: 23159597

114. Doolittle WF. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. Trends Genet. 1998; 14(8):307–311. https://doi.org/10.1016/s0168-9525(98)01494-2 PMID: 9724962.

115. Husnik F, McCutcheon JP. Functional horizontal gene transfer from bacteria to eukaryotes. Nat Rev Microbiol. 2018; 16(2):67–79. Epub 20171127. https://doi.org/10.1038/nrmicro.2017.137 PMID: 29176581.

116. Savory F, Leonard G, Richards TA. The role of horizontal gene transfer in the evolution of the oomycetes. PLoS Pathog. 2015; 11(5):e1004805. https://doi.org/10.1371/journal.ppat.1004805 PMID: 26020232

117. Sibbald SJ, Eme L, Archibald JM, Roger AJ. Lateral gene transfer mechanisms and pan-genomes in eukaryotes. Trends Parasitol. 2020; 36(11):927–41. Epub 20200819. https://doi.org/10.1016/j.pt.2020.07.014 PMID: 32828660.

118. Milner DS, Attah V, Cook E, Maguire F, Savory FR, Morrison M, et al. Environment-dependent fitness gains can be driven by horizontal gene transfer of transporter-encoding genes. Proc Natl Acad Sci U S A. 2019; 116(12):5613–5622. https://doi.org/10.1073/pnas.1815994116 PMID: 30842288

119. Keeling PJ. Horizontal gene transfer in eukaryotes: aligning theory with data. Nat Rev Genet. 2024. Epub 20240123. https://doi.org/10.1038/s41576-023-00688-5 PMID: 38263430.

120. Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD. The origin of plastids. Philos Trans R Soc Lond B Biol Sci. 2008; 363(1504):2675–2685. https://doi.org/10.1098/rstb.2008.0050 PMID: 18468982

121. Gray MW. Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria. Proc Natl Acad Sci U S A. 2015; 112(33):10133–10138. https://doi.org/10.1073/pnas.1421379112 PMID: 25848019

122. Ku C, Nelson-Sathi S, Roettger M, Garg S, Hazkani-Covo E, Martin WF. Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. Proc Natl Acad Sci U S A. 2015; 112(33):10139–10146. https://doi.org/10.1073/pnas.1421385112 PMID: 25733873

123. Esser C, Martin W, Dagan T. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 2007; 3(2):180–184. https://doi.org/10.1098/rsbl.2006.0582 PMID: 17251118; PubMed Central PMCID: PMC2375920.

124. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, et al. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. Cell. 2010; 140(5):631–642. https://doi.org/10.1016/j.cell.2010.01.032 PMID: 20211133

125. Hartman H, Fedorov A. The origin of the eukaryotic cell: a genomic investigation. Proc Natl Acad Sci U S A. 2002; 99(3):1420–1425. https://doi.org/10.1073/pnas.032658599 PMID: 11805300

126. Jian H, Lesley JC. Eukaryotic signature proteins. J Proteom Genom Res. 2012; 1(1):2–8. https://doi.org/10.14302/issn.2326-0793.jpgr-12-101

127. Koonin EV, Yutin N. The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. Cold Spring Harb Perspect Biol. 2014; 6(4):a016188. Epub 20140401. https://doi.org/10.1101/cshperspect.a016188 PMID: 24691961; PubMed Central PMCID: PMC3970416.

128. Baños H, Susko E, Roger AJ. Is over-parameterization a problem for profile mixture models? Syst Biol. 2023:syad063. https://doi.org/10.1093/sysbio/syad063 PMID: 37843172

129. 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 PMID: 29688541

130. 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 PMID: 17652333

131. 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 PMID: 31688943

132. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLoS Biol. 2020; 18(11):e3000862. https://doi.org/10.1371/journal.pbio.3000862 PMID: 33137085

133. Ohno S. Evolution by gene duplication. Springer-Verlag; 1970.

134. Sebé-Pedrós A, Grau-Bové X, Richards TA, Ruiz-Trillo I. Evolution and classification of myosins, a paneukaryotic whole-genome approach. Genome Biol Evol. 2014; 6(2):290–305. https://doi.org/10.1093/gbe/evu013 PMID: 24443438; PubMed Central PMCID: PMC3942036.

135. Diekmann Y, Seixas E, Gouw M, Tavares-Cadete F, Seabra MC, Pereira-Leal JB. Thousands of Rab GTPases for the cell biologist. PLoS Comput Biol. 2011; 7(10):e1002217. https://doi.org/10.1371/journal.pcbi.1002217 PMID: 22022256

136. Jackson CL, Ménétrey J, Sivia M, Dacks JB, Eliáš M. An evolutionary perspective on Arf family GTPases. Curr Opin Cell Biol. 2023; 85:102268. https://doi.org/10.1016/j.ceb.2023.102268 PMID: 39491309

137. van Wijk LM, Snel B. The first eukaryotic kinome tree illuminates the dynamic history of present-day kinases. bioRxiv. 2020:2020.01.27.920793. https://doi.org/10.1101/2020.01.27.920793

138. Archibald JM, Logsdon JM Jr, Doolittle WF. Origin and evolution of eukaryotic chaperonins: phylogenetic evidence for ancient duplications in CCT genes. Mol Biol Evol. 2000; 17(10):1456–1466. https://doi.org/10.1093/oxfordjournals.molbev.a026246 PMID: 11018153.

139. 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 PMID: 30429611

140. Eglit Y, Shiratori T, Jerlström-Hultqvist J, Williamson K, Roger AJ, Ishida K-I, et al. *Meteora sporadica*, a protist with incredible cell architecture, is related to Hemimastigophora. bioRxiv. 2023:2023.08.13.553137. https://doi.org/10.1101/2023.08.13.553137

141. Tikhonenkov DV, Mikhailov KV, Gawryluk RMR, Belyaev AO, Mathur V, Karpov SA, 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 PMID: 36477531

142. Lewin HA, Richards S, Lieberman Aiden E, Allende ML, Archibald JM, Bálint M, et al. The Earth BioGenome Project 2020: Starting the clock. Proc Natl Acad Sci U S A. 2022; 119(4). https://doi.org/10.1073/pnas.2115635118 PMID: 35042800; PubMed Central PMCID: PMC8795548.

143. Consortium DToLP. Sequence locally, think globally: the Darwin tree of life project. Proc Natl Acad Sci U S A. 2022; 119(4). https://doi.org/10.1073/pnas.2115642118 PMID: 35042805; PubMed Central PMCID: PMC8797607.

144. McKenna V, Archibald JM, Beinart R, Dawson M, Hentschel U, Keeling PJ, et al. The aquatic symbiosis genomics project: probing the evolution of symbiosis across the tree of life. Wellcome Open Res. 2021; 6:254. https://doi.org/10.12688/wellcomeopenres.17222.1

145. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016; 1:16048. Epub 20160411. https://doi.org/10.1038/nmicrobiol.2016.48 PMID: 27572647.

146. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008; 9(1):386. https://doi.org/10.1186/1471-2105-9-386 PMID: 18803844

147. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2020; 36(6):1925–1927. https://doi.org/10.1093/bioinformatics/btz848 PMID: 31730192

148. Szánthó LL, Lartillot N, Szöllősi GJ, Schrempf D. Compositionally constrained sites drive long-branch attraction. Syst Biol. 2023; 72(4):767–780. https://doi.org/10.1093/sysbio/syad013 PMID: 36946562

149. Minh BQ, Dang CC, Vinh LS, Lanfear R. QMaker: fast and accurate method to estimate empirical models of protein evolution. Syst Biol. 2021; 70(5):1046–1060. https://doi.org/10.1093/sysbio/syab010 PMID: 33616668

150. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

151. Moi D, Bernard C, Steinegger M, Nevers Y, Langleib M, Dessimoz C. Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses. bioRxiv. 2023:2023.09.19.558401. https://doi.org/10.1101/2023.09.19.558401

152. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet. 2005; 39:309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725 PMID: 16285863.

153. Yan K-K, Wang D, Rozowsky J, Zheng H, Cheng C, Gerstein M. OrthoClust: an orthology-based network framework for clustering data across multiple species. Genome Biol Evol. 2014; 15(8):R100. https://doi.org/10.1186/gb-2014-15-8-r100 PMID: 25249401

154. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13(9):2178–2189. https://doi.org/10.1101/gr.1224503 PMID: 12952885; PubMed Central PMCID: PMC403725.

155. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019; 20(1):238. Epub 20191114. https://doi.org/10.1186/s13059-019-1832-y PMID: 31727128; PubMed Central PMCID: PMC6857279.

156. Liebeskind BJ, McWhite CD, Marcotte EM. Towards consensus gene ages. Genome Biol Evol. 2016; 8(6):1812–1823. https://doi.org/10.1093/gbe/evw113 PMID: 27259914

157. Deutekom ES, Snel B, van Dam TJP. Benchmarking orthology methods using phylogenetic patterns defined at the base of eukaryotes. Brief Bioinform. 2021; 22(3):bbaa206. https://doi.org/10.1093/bib/bbaa206 PMID: 32935832

158. Elias M, Brighouse A, Gabernet-Castello C, Field MC, Dacks JB. Sculpting the endomembrane system in deep time: high resolution phylogenetics of Rab GTPases. J Cell Sci. 2012; 125(10):2500–2508. https://doi.org/10.1242/jcs.101378 PMID: 22366452

159. Vargová R, Wideman JG, Derelle R, Klimeš V, Kahn RA, Dacks JB, et al. A eukaryote-wide perspective on the diversity and evolution of the ARF GTPase protein family. Genome Biol Evol. 2021; 13(8): evab157. https://doi.org/10.1093/gbe/evab157 PMID: 34247240

160. Salomaki ED, Eme L, Brown MW, Kolisko M. Releasing uncurated datasets is essential for reproducible phylogenomics. Nat Ecol Evol. 2020; 4(11):1435–1437. https://doi.org/10.1038/s41559-020-01296-w PMID: 32884150

161. Holland BR, Ketelaar-Jones S, O'Mara AR, Woodhams MD, Jordan GJ. Accuracy of ancestral state reconstruction for non-neutral traits. Sci Rep. 2020; 10(1):7644. https://doi.org/10.1038/s41598-020-64647-4 PMID: 32376845

162. Gàlvez-Morante A, Guéguen L, Natsidis P, Telford MJ, Richter DJ. Dollo parsimony overestimates ancestral gene content reconstructions. Genome Biol Evol. 2024; 16(4):evae062. https://doi.org/10.1093/gbe/evae062 PMID: 38518756

163. Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. Efficient exploration of the space of reconciled gene trees. Syst Biol. 2013; 62(6):901–912. https://doi.org/10.1093/sysbio/syt054 PMID: 23925510

164. Derelle R, Torruella G, Klimeš V, Brinkmann H, Kim E, Vlček Č, et al. Bacterial proteins pinpoint a single eukaryotic root. Proc Natl Acad Sci U S A. 2015; 112(7):E693–E699. https://doi.org/10.1073/pnas.1420657112 PMID: 25646484

165. He D, Fiz-Palacios O, Fu CJ, Fehling J, Tsai CC, Baldauf SL. An alternative root for the eukaryote tree of life. Curr Biol. 2014; 24(4):465–70. Epub 20140206. https://doi.org/10.1016/j.cub.2014.01.036 PMID: 24508168.

166. Al Jewari C, Baldauf SL. An excavate root for the eukaryote tree of life. Sci Adv. 2023; 9(17): eade4973. Epub 20230428. https://doi.org/10.1126/sciadv.ade4973 PMID: 37115919; PubMed Central PMCID: PMC10146883.

167. Roger AJ, Williamson K, Eme L, Baños H, McCarthy C, Susko E, et al. A robustly rooted tree of eukaryotes reveals their excavate ancestry. Res Sq. 2024. https://doi.org/10.21203/rs.3.rs-5059906/v1

168. Nevers Y, Jones TEM, Jyothi D, Yates B, Ferret M, Portell-Silva L, et al. The Quest for Orthologs orthology benchmark service in 2022. Nucleic Acids Res. 2022; 50(W1):W623–W632. https://doi.org/10.1093/nar/gkac330 PMID: 35552456

169. Dessimoz C, Gabaldón T, Roos DS, Sonnhammer ELL, Herrero J, Consortium tQfO. Toward community standards in the quest for orthologs. Bioinformatics. 2012; 28(6):900–904. https://doi.org/10.1093/bioinformatics/bts050 PMID: 22332236

170. Gough J, Karplus K, Hughey R, Chothia C. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol. 2001; 313(4):903–919. https://doi.org/10.1006/jmbi.2001.5080 PMID: 11697912.

171. Eddy SR. Hidden Markov models. Curr Opin Struct Biol. 1996; 6(3):361–365. https://doi.org/10.1016/s0959-440x(96)80056-x PMID: 8804822

172. Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2005; 21(7):951–960. https://doi.org/10.1093/bioinformatics/bti125 PMID: 15531603

173. Jablonowski K. Hidden Markov Models for protein domain homology Identification and analysis. Methods Mol Biol. 2017; 1555:47–58. https://doi.org/10.1007/978-1-4939-6762-9_3 PMID: 28092026.

174. Srivastava PK, Desai DK, Nandi S, Lynn AM. HMM-ModE—improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. BMC Bioinformatics. 2007; 8:104. Epub 20070327. https://doi.org/10.1186/1471-2105-8-104 PMID: 17389042; PubMed Central PMCID: PMC1852395.

175. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016; 44(D1):D457–D462. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

176. Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, et al. 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 PMID: 39431411

177. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol. 2019; 1962:227–245. https://doi.org/10.1007/978-1-4939-9173-0_14 PMID: 31020564.

178. Nevers Y, Warwick Vesztrocy A, Rossier V, Train C-M, Altenhoff A, Dessimoz C, et al. Quality assessment of gene repertoire annotations with OMArk. Nat Biotechnol. 2024. https://doi.org/10.1038/s41587-024-02147-w PMID: 38383603

179. Tice AK, Žihala D, Pánek T, Jones RE, Salomaki ED, Nenarokov S, et al. PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. PLoS Biol. 2021; 19(8):e3001365. https://doi.org/10.1371/journal.pbio.3001365 PMID: 34358228

180. Archibald JM. The puzzle of plastid evolution. Curr Biol. 2009; 19(2):R81–R88. https://doi.org/10.1016/j.cub.2008.11.067 PMID: 19174147

181. Keeling PJ. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol. 2013; 64(1):583–607. https://doi.org/10.1146/annurev-arplant-050312-120144 PMID: 23451781

182. Mathur V, Kolísko M, Hehenberger E, Irwin NAT, Leander BS, Kristmundsson Á, et al. Multiple independent origins of apicomplexan-like parasites. Curr Biol. 2019; 29(17):2936–41.e5. https://doi.org/10.1016/j.cub.2019.07.019 PMID: 31422883

183. Richards TA, Soanes DM, Jones MDM, Vasieva O, Leonard G, Paszkiewicz K, et al. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci U S A. 2011; 108(37):15258–15263. https://doi.org/10.1073/pnas.1105100108 PMID: 21878562

184. Bartošová-Sojková P, Butenko A, Richtová J, Fiala I, Oborník M, Lukeš J. Inside the host: understanding the evolutionary trajectories of intracellular parasitism. Annu Rev Microbiol. 2024. Epub 20240429. https://doi.org/10.1146/annurev-micro-041222-025305 PMID: 38684082.

185. Poulin R, Randhawa HS. Evolution of parasitism along convergent lines: from ecology to genomics. Parasitology. 2015; 142(Suppl 1):S6–s15. Epub 20131111. https://doi.org/10.1017/S0031182013001674 PMID: 24229807; PubMed Central PMCID: PMC4413784.

186. Herron MD, Conlin PL, Ratcliff WC. The evolution of multicellularity. Taylor & Francis Group; 2022.

187. Blaz J, Galindo LJ, Heiss AA, Kaur H, Torruella G, Yang A, et al. One high quality genome and two transcriptome datasets for new species of *Mantamonas*, a deep-branching eukaryote clade. Sci Data. 2023; 10(1):603. https://doi.org/10.1038/s41597-023-02488-2 PMID: 37689692

188. Horváthová L, Žárský V, Pánek T, Derelle R, Pyrih J, Motyčková A, et al. Analysis of diverse eukaryotes suggests the existence of an ancestral mitochondrial apparatus derived from the bacterial type II secretion system. Nat Commun. 2021; 12(1):2947. https://doi.org/10.1038/s41467-021-23046-7 PMID: 34011950

189. Cotton JA, McInerney JO. Eukaryotic genes of archaebacterial origin are more important than the more numerous eubacterial genes, irrespective of function. Proc Natl Acad Sci U S A. 2010; 107(40):17252–5. Epub 20100917. https://doi.org/10.1073/pnas.1000265107 PMID: 20852068; PubMed Central PMCID: PMC2951413.