nature ecology & evolution

Article

https://doi.org/10.1038/s41559-025-02648-0

Chimeric origins and dynamic evolution of central carbon metabolism in eukaryotes

Received: 29 May 2024

Accepted: 24 January 2025

Published online: 3 March 2025

Check for updates

Carlos Santana-Molina [1], Tom A. Williams [2], Berend Snel [3,5] & Anja Spang [1,4,5]

The origin of eukaryotes was a key event in the history of life. Current leading hypotheses propose that a symbiosis between an asgardarchaeal host cell and an alphaproteobacterial endosymbiont represented a crucial step in eukaryotic origin and that metabolic cross-feeding between the partners provided the basis for their subsequent evolutionary integration. A major unanswered question is whether the metabolism of modern eukaryotes bears any vestige of this ancestral syntrophy. Here we systematically analyse the evolutionary origins of the eukaryotic gene repertoires mediating central carbon metabolism. Our phylogenetic and sequence analyses reveal that this gene repertoire is chimeric, with ancestral contributions from Asgardarchaeota and Alphaproteobacteria operating predominantly in glycolysis and the tricarboxylic acid cycle, respectively. Our analyses also reveal the extent to which this ancestral metabolic interplay has been remodelled via gene loss, transfer and subcellular retargeting in the >2 billion years since the origin of eukaryotic cells, and we identify genetic contributions from other prokaryotic sources in addition to the asgardarchaeal host and alphaproteobacterial endosymbiont. Our work demonstrates that, in contrast to previous assumptions, modern eukaryotic metabolism preserves information about the nature of the original asgardarchaeal–alphaproteobacterial interactions and supports syntrophy scenarios for the origin of the eukaryotic cell.

The origin of eukaryotes represents a defining event in the history of life that occurred between 2.6 and 1.2 billion years ago (Ga)[1–5], possibly coinciding with rising oxygen levels in the atmosphere of the Earth[6–8]. One of the key steps during eukaryogenesis involved symbiosis between a member of the Asgardarchaeota[9–12] and a bacterial partner related to the Alphaproteobacteria that evolved to become the mitochondrion[13,14].

To explain the evolutionary driving forces underlying eukaryogenesis, many models have been proposed[15–29] that differ with respect to the identity and number of partners involved and the nature of their initial interactions, ranging from syntrophy to phagocytosis and parasitism. The discovery of the Asgardarchaeota lent support to hypotheses invoking a syntrophic relationship between at least one archaeal and bacterial partner[26,28], and metabolic capabilities inferred for the asgardarchaeal ancestor of eukaryotes have inspired updated hypotheses about the syntrophic interactions between the archaeal and bacterial partners during the early stages of eukaryogenesis[24,25,30]. These syntrophy-based hypotheses suggest that one partner may have been dependent on the other as an external electron sink, but

[1]Department of Marine Microbiology and Biogeochemistry, NIOZ, Royal Netherlands Institute for Sea Research, AB Den Burg, the Netherlands. [2]Bristol Palaeobiology Group, School of Biological Sciences, University of Bristol, Bristol, UK. [3]Theoretical Biology & Bioinformatics, Department of Biology, Faculty of Science, Utrecht University, Utrecht, the Netherlands. [4]Department of Evolutionary & Population Biology, Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, the Netherlands. [5]These authors contributed equally: Berend Snel, Anja Spang. e-mail: b.snel@uu.nl; anja.spang@nioz.nl

make distinct predictions about the types of metabolites exchanged, the origin of eukaryotic cell membranes, the timing of mitochondrial acquisition, the mechanism of mitochondrial uptake and the origin of the nucleus[26,28,31–33]. However, testing these hypotheses with current data is challenging, in part because the evolutionary origin of eukaryotic metabolism remains understudied.

Previous genomic analyses have suggested that 'informational' genes (those involved in translation, replication and transcription) generally have archaeal origins, while 'operational' genes (those involved in metabolism) derive predominantly from bacteria, particularly from the premitochondrial endosymbiont[34–39]. This has led to the hypothesis that, during eukaryogenesis, the archaeal host metabolism was replaced by counterparts from the endosymbiont[15]. However, considering that syntrophy relies on metabolic repertoires from both partners, archaeal gene contributions to eukaryotic metabolisms might be expected.

To assess current models on the origins of eukaryotic cells and the evolution of plastids, we here analyse the origins of eukaryotic central carbon metabolism (CCM), comprising four main pathways: the Embden–Meyerhof–Parnas (EMP) and the Entner–Doudoroff glycolytic pathways, the pentose phosphate pathway (PPP), the pyruvate/acetate conversions into acetyl-CoA and the tricarboxylic acid (TCA) cycle (Supplementary Discussion). While it has previously been suggested that many TCA cycle enzymes, as well as enzymes involved in pyruvate conversions, were present in last eukaryotic common ancestor (LECA) and trace their origin back to Alphaproteobacteria[13,40–42], the evolutionary origins of glycolysis and PPP in eukaryotes remain unresolved[20,40]. Furthermore, several genes involved in eukaryotic metabolism appear to have origins unrelated to either symbiotic partner, potentially reflecting independent horizontal gene transfer (HGT) acquisitions either before or after the radiation of the extant eukaryotic lineages[43–45], further complicating phylogenetic analyses. The metagenomics-based discovery of new archaeal and bacterial lineages during the past decades, including the Asgardarchaeota, has provided a wealth of new information to address the origins and evolution of eukaryotic CCM within the context of a more broadly sampled tree of life[9,46–49]. Our comprehensive phylogenetic analyses reveal a much more complex pattern of evolution than previously anticipated and identify a chimeric CCM that includes contributions of archaeal origin to the LECA proteome. The distribution of CCM enzymes across the eukaryotic tree of life illuminates the subsequent highly dynamic evolution of these enzyme repertoires shaped by gene loss, endosymbiotic gene transfers (EGTs) and gene replacements.

Results

Central carbon metabolism of LECA

We selected a balanced and representative set of 207 eukaryotic proteomes that cover currently known taxonomic diversity and lifestyles, including anaerobic eukaryotes and eukaryotes with primary and higher-order plastid organelles (Fig. 1a and Supplementary Data 1). To compare gene trees of CCM enzymes with the eukaryotic tree of life, we first reconstructed a species tree on the basis of a manually curated set of 317 concatenated phylogenetic markers[50]. We used both maximum-likelihood and Bayesian approaches combined with trimming of heterogeneous sites to evaluate the robustness of support for major eukaryotic clades (Methods; Fig. 1, Extended Data Fig. 1 and Supplementary Discussion). The resulting tree comprises three major supergroups (Fig. 1a): Excavata, Amorphea and Diaphoretickes. Although we rooted the tree between Excavata and the other groups for visualization, the placement of the root remains under debate[48,51–55] and our interpretations of gene family origins do not assume a particular root position. Excavata include Jakobids (within Discoba) which, together with Mantamonas (within CRuMs as part of Amorphea), possess one of the most gene-rich mitogenomes among eukaryotes[56–60]. Another lineage of the Excavata is Metamonada, members of which are anaerobic and contain mitochondrial-related organelles that have been entirely lost in some representatives[61–64]. Metamonada often form long branches in phylogenetic trees, which hampers the phylogenetic placement of some putative member lineages such as the Anaeramoebae[65], which in our analyses alternatively branch with Amorphea clades under a subset of tree inference parameters (Extended Data Fig. 1b,c). Amorphea include Obozoa (Fungi, Metazoa and various protists), Amoebozoa and other putative taxa such as CRuMs, Malawimonada and Ancyromonada. The monophyly of Amorphea is not fully stable: specifically, while Ancyromonadida most consistently places within Amorphea, we also observed its clustering sister to Diaphoretickes (Extended Data Fig. 1c and Supplementary Discussion). Diaphoretickes form a stable group composed of two well-supported monophyletic clades: the Cryptista–Archaeplastida (the last including Chloroplastida, Rhodophyta, Glaucophyta, Picozoa and Rhodelphis) and SAR (Stramenopiles, Alveolata and Rhizaria). Apart from these, Haptista, Telonemia, Hemimastigophora and Anconracysta (the last now classified as Provora phylum[66]) showed varying placements within Diaphoretickes, depending on site filtering and phylogenetic methods (Extended Data Fig. 1). Overall, the inferred eukaryotic tree of life, and the inference of the three major supergroups, Excavata, Amorphea and Diaphoretickes, provide a solid framework for interpreting individual gene trees, defining LECA versus post-LECA clades and thereby determining the relative timing of gene acquisitions.

Next, we inferred the phylogenies of 64 gene families encoding enzymes involved in the CCM of eukaryotes and evaluated the evolutionary origins of eukaryotic homologues in each phylogeny (Supplementary Figs. 1–32 and Supplementary Discussion). CCM enzyme gene family membership was determined using protein model annotations based on the KEGG orthology database (Supplementary Data 2) providing the starting point for the collection of homologues for phylogenetic inferences. Phylogenetic analyses were performed iteratively to maximize resolution and flag problematic data as well as putative eukaryotic contaminations (Extended Data Fig. 2 and Supplementary Data 3; Methods). For each CCM enzyme family, we manually identified putative ancestral clades in eukaryotes, including those containing organisms from at least two major groups, Excavata, Amoprhea and Diaphoretickes (that is, potential LECA clades). The distribution of these orthogroups was mapped onto the eukaryotic species tree (Fig. 1), showing the widespread distribution of these enzymes in most eukaryotic groups and suggesting the presence of a canonical CCM in LECA.

Specifically, our phylogenetic analyses suggest that nine out of ten EMP glycolysis (Fig. 1 and Supplementary Figs. 1–12) and seven out of eight PPP (Supplementary Figs. 13–22 and Supplementary Discussion) enzymatic steps comprise putative LECA clades. By contrast, phylogenies of Entner–Doudoroff glycolytic enzymes seem to indicate that these enzymes have been acquired later during eukaryotic evolution in photosynthetic eukaryotes and representatives with secondary endosymbionts[67] (Supplementary Fig. 21). For pyruvate/acetate conversions, we observed that the pyruvate dehydrogenase complex (PDH, formed by PDHA/B/C/D subunits) as well as acetyl-CoA synthetase (ACS) and lactate dehydrogenases (LDH), were present in LECA. In contrast, the respective analogous enzymes for these reactions—pyruvate formate lyase, pyruvate-ferredoxin/flavodoxin oxidoreductase (POR) and ADP-forming acetyl-CoA synthetase (ACDA/B)—were found in extant anaerobic Metamonada, Archamoebae and Breviatea and some aerobic organisms (Fig. 1, Supplementary Figs. 23–27 and Supplementary Discussion). Most of these later cases probably reflect post-LECA acquisitions with subsequent transfer among eukaryotes, while others (such as POR) may have been present in LECA[45]. The phylogenies of the reverse TCA cycle defined by ATP-citrate lyase subunits (ACLA/B) or its fused version (ACLY; Fig. 1 and Supplementary Fig. 28), suggested that both were probably present in LECA as proposed previously[68]. Finally, for the TCA cycle, all phylogenies for the ten enzymatic steps

https://doi.org/10.1038/s41559-025-02648-0

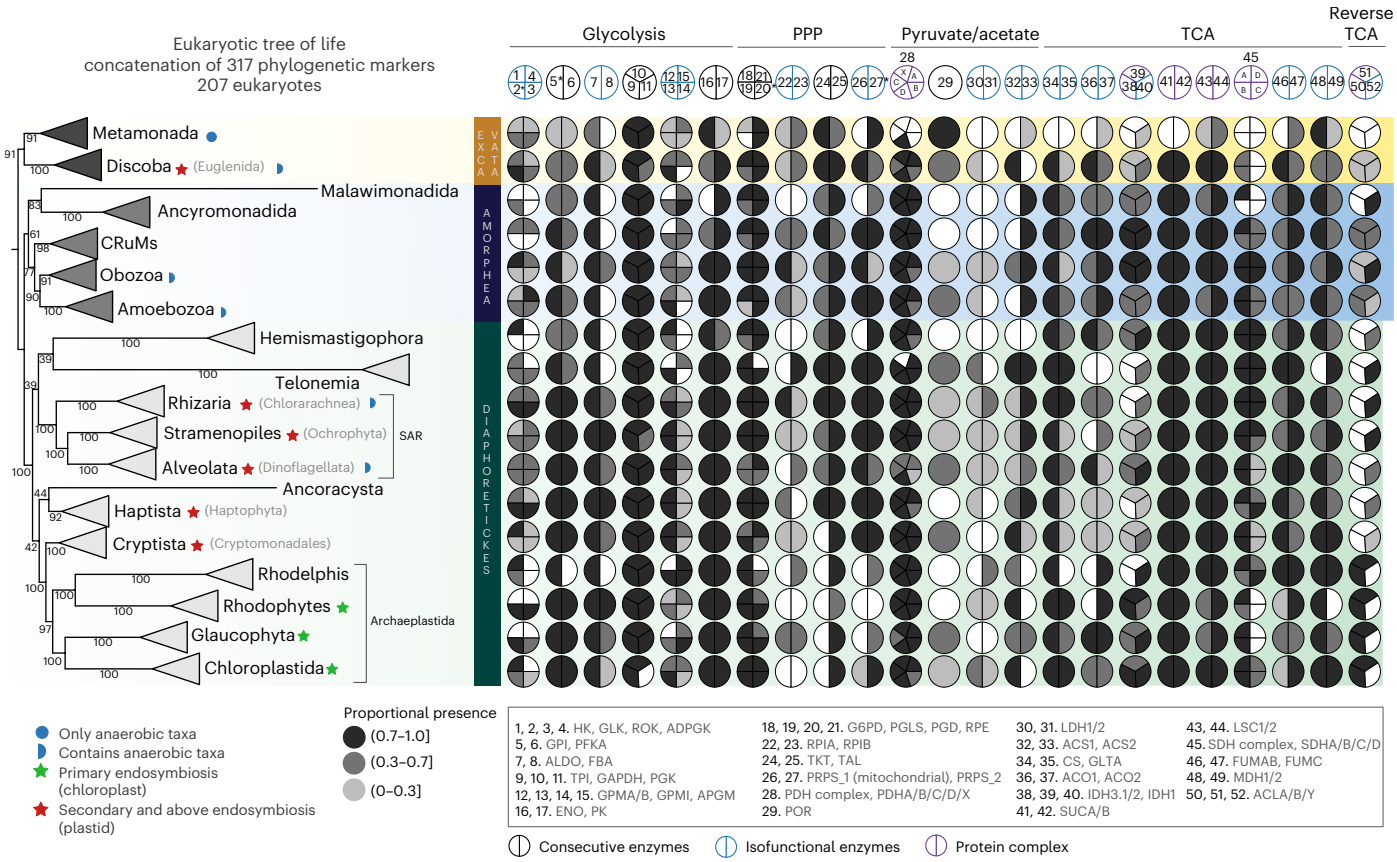


Fig. 1 | Phylogenetic reconstruction of the eukaryotic tree of life. Left, maximum-likelihood phylogeny of the eukaryotic tree of life based on the concatenation of 317 phylogenetic markers. The tree is unrooted, but drawn with Excavata at the root for ease of visualization. The concatenated multiple sequence alignment (MSA) consisted of 207 taxa, 97,680 positions and the tree was built with IQ-TREE.2.1.2 under the LG + C60 + G model and using optimized ultrafast bootstrap (Ufboot2 -bnni). Annotation corresponds to characteristic traits (see legend). Extended tree and additional phylogenetic analyses of eukaryotic species tree provided in Extended Data Fig. 1. Right, global phylogenetic distribution of CCM enzymes across eukaryotic species trees. HK, hexokinase; GLK, glucokinase; ROK, repressor protein, open reading frame, sugar kinase (that is, hexokinase); ADPGK, ADP-dependent glucokinase; GPI, glucose-6-phosphate isomerase; PFKA, 6-phosphofructose kinase (A); ALDO, fructose-biphosphate aldolase, Class 1; FBA, fructose-biphosphate aldolase; TPI, triosephosphate isomerase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; PGK, phosphoglycerate kinase; GPMA/GPMB/GPMI/APGM, 2,3-bisphosphoglycerate phosphoglycerate mutases; ENO, enolase; PK, pyruvate kinase; G6PD, glucose 6-phosphate dehydrogenase; PGLS, 6-phosphogluconolactonase; PGL, 6-phosphogluconolactonase; PGD, 6-phosphogluconate dehydrogenase; RPE, ribulose-phosphate 3-epimerase; RPIA/B, ribose 5-phosphate isomerase A/B; TKT, transketalase; TAL, transaldolase; PRPS, ribose-phosphate pyrophosphokinase; PDH, pyruvate dehydrogenase complex; POR, pyruvate-ferredoxin/flavodoxin oxidoreductase; LDH, lactate dehydrogenase; ACS, acetyl-CoA synthetase; CS/GLTA, citrate synthase; ACO, aconitate hydratase; IDH, isocitrate dehydrogenase; SUC (A/B), 2-oxoglutarate dehydrogenase complex, subunit A/B, succinyl-CoA synthetase complex; SDH (A/B/C/D), succinate dehydrogenase complex (A/B/C/D); FUM (AB/C), fumarate hydratases (AB/C); MDH, malate dehydrogenase; ACL, ATP-citrate lyase (see also Supplementary Data 2). Pie charts group consecutive, isofunctional and protein complex enzymes and different grey shading indicates the proportion of taxa from the respective taxonomic level bearing such gene. Orthogroups were manually selected from phylogenetic trees. Asterisks denote those trees containing paraphyletic clades with unclear origins (Supplementary Discussion).

showed clear LECA clades (Fig. 1, Supplementary Fig. 29–37 and Supplementary Discussion), indicating that the TCA was present in LECA. Yet, the TCA was nearly absent in Metamonada, Archamoebae, Microsporidia and partially in Breviatea, in line with suggested secondary losses in these eukaryotic clades[64]. As previously observed[64,65,69,70], mitochondrial-derived PDHD and fumarate dehydrogenase C (FUMC, see below) of metamonads branch within LECA clades (Fig. 1 and Supplementary Figs. 23 and 35b), consistent with the prevailing view that these organisms secondarily lost mitochondria, rather than that they never had them[55]. Notably, the phylogeny of some enzymatic steps associated with these metabolic pathways revealed the presence of independent orthogroups coding for enzymes predicted to perform the same reactions. Examples include phosphoglycerate mutases (in EMP), citrate synthase (CS), aconitases (ACO) and isocitrate dehydrogenases (IDH, in TCA), suggesting that LECA harboured metabolic redundancy for these metabolic steps (Fig. 1 and Supplementary Figs. 9, 10, 29, 30 and 31). Altogether, these phylogenies unambiguously show that LECA had the complete set of enzymes needed for the CCM.

Prokaryotic origins of eukaryotic CCM

Next, we inferred the prokaryotic donor lineages for CCM enzymes present in LECA (summarized in Fig. 2a and detailed origins and distribution in Fig. 2b, Extended Data Fig. 3 and Supplementary Data 4), classifying gene trees by inferred donor lineage including Asgardarchaeota, the mitochondrial and chloroplast endosymbionts and other prokaryotic taxa. In what follows, we discuss examples of particular interest for each category. The associated phylogenetic trees are presented in Fig. 3a–d and Supplementary Figs. 1–37, with Fig. 3e depicting the general taxonomic composition of prokaryotic sister groups to enzymes present in LECA.

Asgardarchaeal host contributions. In four EMP phylogenies, putative LECA clades branch sister to Asgardarchaeota (Figs. 2 and 3a and Supplementary Figs. 4, 12 and 13): ADP-dependent glucokinase (ADPGK, acting in the first and third step of EMP[71]), two 2,3-bisphosphoglycerate-independent mutases (APGM and GPMI, analogous enzymes, see also Supplementary Discussion) and enolase

https://doi.org/10.1038/s41559-025-02648-0

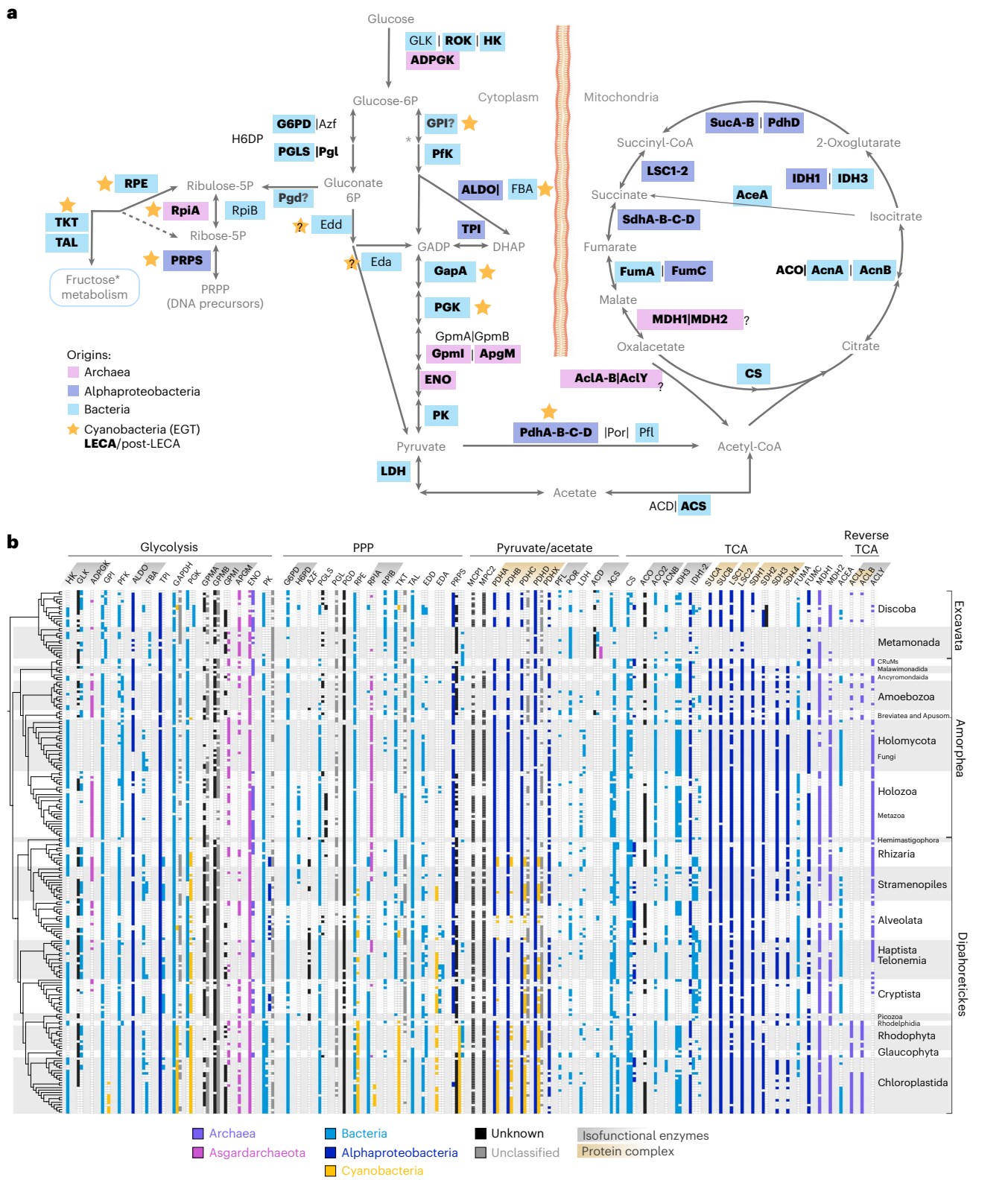


Fig. 2 | Prokaryotic origins and distribution of eukaryotic CCM. a, CCM pathways highlighting the proposed origins by colours. Enzyme names (see Fig. 1, Supplementary Discussion and Supplementary Data 2) in bold denote those enzymes potentially present in LECA. Edd, phosphogluconate dehydratase, dihydroxy-acid dehydratase; Eda, 2-dehydro-3-deoxyphosphogluconate aldolase; Azf, NAD^+ dependent glucose-6-phosphate dehydrogenase; Acd, acetate–CoA ligase; AceA, isocitrate lyase; Pfl, formate C-acetyltransferase (Supplementary Data 2). Asterisks indicate tentative inferences. **b**, Phylogenetic profile of eukaryotic orthogroups manually selected from phylogenetic trees. Coloured cells indicate presence of genes with a proposed origin, see legend. Unknown origins (black cells), refer to those unresolved phylogenies. Grey cells column indicate sequences not considered as orthogroups (unclassified) and they are shown if they are present in >80 eukaryotes. Bold enzyme names are those hypothesized to be present in LECA and grey and light-brown denote isofunctional and protein complex enzymes, respectively. Eukaryotic tree of life includes fast-evolving and low genome completeness taxa such as Microsporidia and Picozoa (Extended Data Fig. 3). Raw data for presence/absence profile provided in Supplementary Data 4.

(ENO). In three of these phylogenies, eukaryotes are nested within Archaea sister to sequences from Asgardarchaeota, consistent with the putative inheritance of these enzymes from the asgardarchaeal ancestor of eukaryotes (ADPGK 76%, APGM 89% and ENO 39% UfBoot2). In contrast, the topology of GPMI_1 shows a sister relationship between a few asgardarchaeal and eukaryotic sequences. In the PPP, the ribose 5-phosphate isomerase A (RPIA) phylogeny shows a large eukaryotic cluster containing species from the Amorphea and few Discoba branching next to Asgardarchaeota (97%; Figs. 2 and 3a and Supplementary Fig. 16). Together, this may indicate that Asgardarchaeota have contributed gene families to the CCM of eukaryotes.

Other cases of potential archaeal origins remain more speculative. Eukaryotic ATP-citrate lyase subunits (ACLA/B) or its fused version (ACLY), appear to be present in various Asgardarchaeota and might therefore have been inherited by eukaryotes through the host lineage. However, the eukaryotic ACLA/B and ACLY clades branch with DPANN[72,73] and Thermoplasmatota-E3, respectively, suggesting independent origins from different archaeal groups (Supplementary Fig. 28 and Supplementary Discussion). Eukaryotic malate dehydrogenase LECA paralogues operate in the cytoplasm (MDH1) and in the mitochondria (MDH2)[74]. The phylogeny of the MDH family in combination with conserved spliceosomal intron positions (Malin[75] LECA intron probability 0.6; Methods) suggest that MDH1 and MDH2 originated by duplication before LECA, with a clade containing TACK[76] archaeal sequences and Baldarchaeota sequences as sister groups (Supplementary Fig. 36, Supplementary Data 5 and Supplementary Discussion). However the long branches characterizing these phylogenetic relationships render the archaeal origin of MDH1/2 tentative.

Lastly, in three phylogenies, we observed the clustering of a limited number of eukaryotes with Asgardarchaeota (Fig. 3d). The pyruvate kinase (PK) phylogeny contains a phylogenetic group, PK_5, mainly composed of Amoebozoa (plus a few others), nested within an asgardarchaeal clade (Fig. 3d and Supplementary Fig. 12d). Furthermore, the phylogeny of GPMI showed, in addition to a clear LECA clade (GMPI_1), a small monophyletic group of Asgardarchaeota with some eukaryotes (labelled as GPMI_2; Fig. 3a, Supplementary Fig. 10 and Supplementary Discussion). The third case is the ACDA/B enzyme family which is involved in the reversible conversion of acetate into acetyl-CoA using ATP (analogue of ACS) (Fig. 2a). We found a fully supported monophyletic group (100% UfBoot2, ACD3) containing Lokiarchaeia and Fornicata, a lineage within Metamonada (Fig. 3d and Supplementary Fig. 27). The clustering of these eukaryotic clades with Asgardarchaeota would be consistent with orthogroups that were present in LECA and subsequently underwent loss during eukaryotic evolution, resulting in their absence from model organisms[77] (that is, vertical gene transfer from Asgardarchaeota to LECA of PK). Alternatively, this clustering might reflect a post-LECA transfer from Asgardarchaeota to the ancestor of one of these eukaryotic groups with subsequent transfer between eukaryotes. Overall, and despite the sometimes limited resolution of single-gene trees, these cases demonstrate a previously underappreciated role of the asgardarchaeal host cell in shaping eukaryotic CCM.

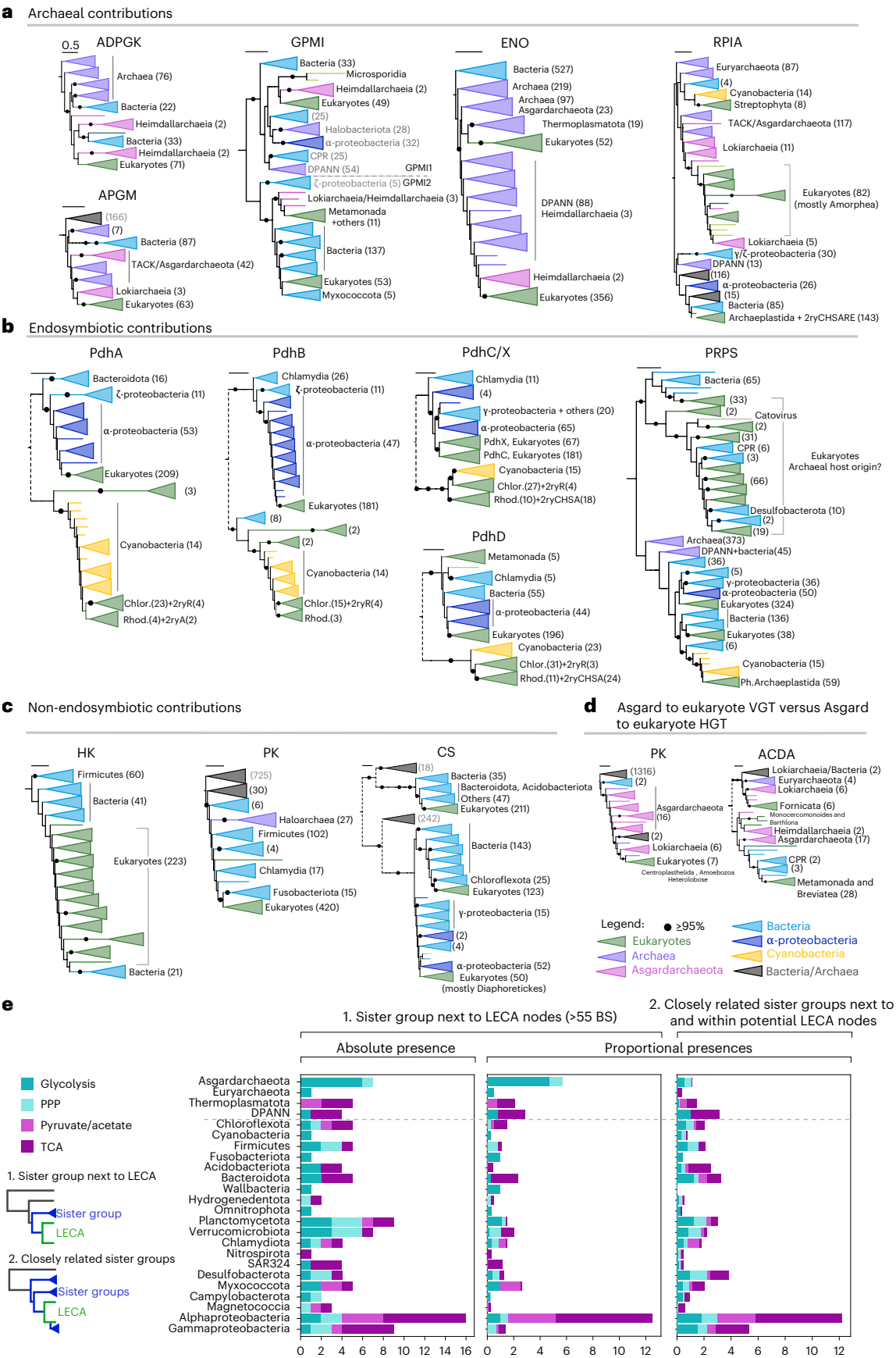
Alphabacterial and cyanobacterial endosymbiotic contributions. The next category involves putative endosymbiotic contributions to LECA and the Archaeplastida ancestor through EGT[78]. Indeed, many CCM phylogenies recovered Alphaproteobacteria as sister clades to eukaryotes (Figs. 2 and 3b). We found two alphaproteobacterial contributions to the EMP (fructose-bisphosphate aldolase (ALDO) and triosephosphate isomerase (TPI)), one to the PPP (ribose phosphate pyrophosphokinase (PRPS)), four in pathways related to pyruvate conversions (PDHA/B/C/D) and ten to the TCA cycle (IDH1, 2-oxoglutarate dehydrogenase subunits, SUCA/B, succinyl-CoA synthetase alpha/beta subunits, LSC1/2, succinate dehydrogenase subunits, SDH1/2/3/4 and FUMC; Supplementary Figs. 5a, 6, 22, 23 and 31–35). Thus, most of the potential contributions from alphaproteobacteria seem to operate in the mitochondria (Fig. 2a).

We also found several likely cyanobacterial contributions to the CCM of photosynthetic eukaryotes (Fig. 2a). Specifically, four enzymes of the EMP (glucose-6-phosphate isomerase (GPI), ALDO class II, FBA, glyceraldehyde 3-phosphate dehydrogenase, (GAPDH) and phosphoglycerate kinase (PGK)), six of the PPP (ribulose-phosphate 3-epimerase (RPE), RPIA, PRPS, transketolase (TKT), phosphogluconate dehydratase (EDD) and 2-dehydro-3-deoxyphosphogluconate aldolase (EDA)) and four among pyruvate conversions (PDHA/B/C/D), have topologies consistent with being derived by EGT from cyanobacteria (Supplementary Figs. 3, 5b, 7, 8 and 17–23). These phylogenetic clusters also contain photosynthetic eukaryotes with higher level plastids (for example, secondary and tertiary endosymbioses). The PDHC/D phylogenies revealed a green algae plastid EGT in Chlorarachnea, as well as red algae plastid EGTs in Cryptophyceae, Haptophyta, Myzozoa and Gyrista (chromist lineage[79]; Fig. 3b and Supplementary Fig. 23). Similarly, the phylogenies of PGK and RPE (Supplementary Figs. 8 and 17) also comprise plastid EGTs, whereas the phylogenies of PFK2, ENO, RPIA, LSC2 and ALDO, showed monophyletic groups comprising red algae derived lineages within LECA clades indicative of nucleus-to-nucleus EGTs (Supplementary Figs. 4, 5, 11 and 32). These observed EGTs involved various sister groups, with eukaryotic taxa ranging from red and green algae to non-photosynthetic organisms and suggesting complex evolutionary histories. Thus, PDHC/D phylogenies support distinct secondary endosymbiosis events involving green and red algae, respectively, and serial endosymbioses in red lineages[50,79].

The phylogenetic signal for the sister relationship between Alphaproteobacteria or Cyanobacteria and eukaryotes is not always unequivocal[13] (Supplementary Discussion): in phylogenies of TPI and PRPS, eukaryotes are sister to Alpha/Gammaproteobacteria, while trees of LSC2 and ACS, recover genes of other prokaryotic clades interspersed between the alphaproteobacterial/LECA clades. Furthermore, the PDHD and FUMC phylogenies include divergent eukaryotic sequences which branch within the Alphaproteobacteria clade rather than within the LECA clade (especially Excavata taxa; Supplementary Figs. 23 and 34). Similarly, cyanobacterial contributions are not always highly supported (for example, RPE, EDD and EDA; Supplementary Figs. 17 and 21). Nevertheless, our analysis shows that alphaproteobacterial contributions

Fig. 3 | Maximum-likelihood phylogenies of selected enzymes representing diverse evolutionary origins of CCM in eukaryotes. a–**d**, Examples of potential archaeal contributions from Asgardarchaeota (**a**), contributions from Alphaproteobacteria and Cyanobacteria (**b**), other prokaryotic origins (from known or unknown donor) (**c**) and Asgardarchaeota to eukaryote vertical gene transfer (VGT) versus HGT (**d**). Discontinuous lines indicate simplified tree topology after pruning the relative branches of interest. Number between brackets denotes the number of sequences in the respective clades. 2ryCHSRA, denotes secondary endosymbionts including Cryptista (C), Haptista (H), Stramenopile (S), Rhizaria (R) and Alveolata (A), Rhodophyta (Rhod), Chlorplastida (Chlor), photosynthetic (Ph) Archaeplastida. Phylogenies were built with IQ-TREE.2.1.2 under the LG + C20 + G + F model and using optimized ultrafast bootstrap (NNI UfBoot2). Extended trees are provided in Supplementary Figs. 1–37. Complete enzyme gene names defined in the text are: ADPGK, APGM and GPMI, ENO, RPIA, PDHA/B/C/D, PRPS, HK, PK, CS and ACDA. **e**, General taxonomic composition of sister group(s) of selected potential LECA clades. First and second panels (1) depict the taxonomic composition of the first sister group to a LECA clade, while the third panel (2) depicts the composition of the LECA closely related sister groups (see legend for visual explanation). Bar length is the sum of the respective presences of a taxon. 'Absolute presence' counts the presence of a specific taxon in the sister group, while 'proportional presences' represent the proportion of a taxon respective to the size of the sister group(s). Those taxa whose proportional presence in the single sister group was ≥1, were shown in the plot. Different coloured stacked bars refer to different pathways and prokaryotic phyla were sorted according to species relationships. BS, bootstrap support.

https://doi.org/10.1038/s41559-025-02648-0



https://doi.org/10.1038/s41559-025-02648-0

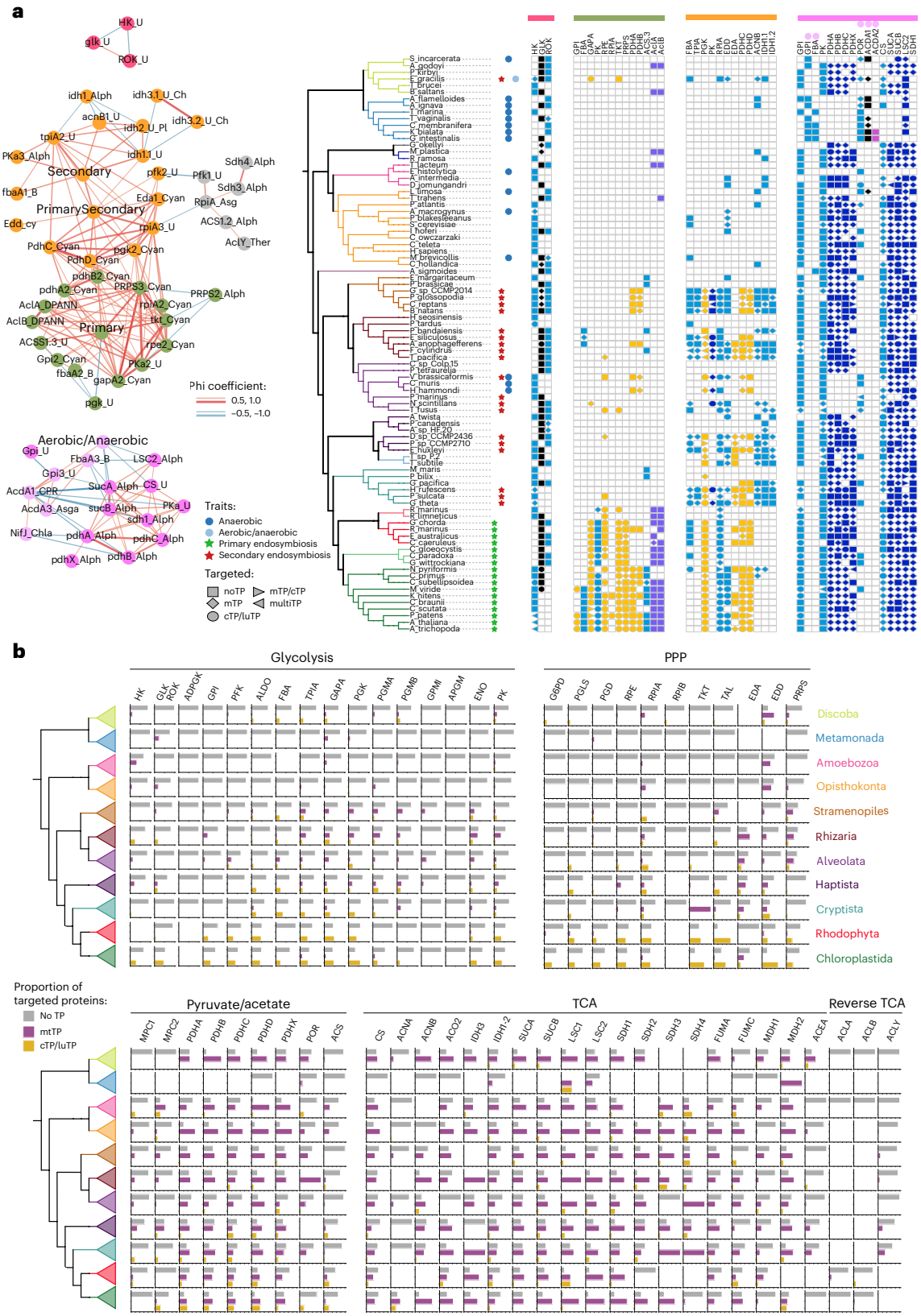


Fig. 4 | Correlative distributions of CCM enzymes across eukaryotes and their targeting in eukaryotic cells. a, Left, correlative networks for the distribution of those orthogroups with higher (red edges) or lower (blue edges) phi coefficients than 0.5 and −0.5, respectively. Light pink indicates orthogroups including anaerobic eukaryotes. Clusters were obtained by modularity using Gephi. Right, phylogenetic profile of the respective correlated orthogroups indicating their evolutionary origins (cell colour) and targeting signal (cell shape). Taxonomic tree is a subselection of representatives and is annotated with characteristic traits of the respective taxa (see legend). **b**, Distribution of targeted proteins along the eukaryotic tree of life and the CCM. Bars represent the proportion of sequences with the respective targeting. noTP, no transit peptide; mTP, mitochondrial transit peptide; cTP, chloroplast transit peptide; luTP, thylakoid luminal transit peptides; multiTP, multiple transit peptides (see legend). LuTPs were clustered together with cTPs. Only sequences from selected orthogroups (Fig. 2b) were used for this analysis. Raw data for these plots provided in Supplementary Data 4.

to LECA mainly operate in the TCA pathway, and that cyanobacterial EGT contributions to the Archaeplastida ancestor often comprise enzymes of the EMP glycolysis and the PPP. The PDH complex and PRPS phylogenies revealed contributions derived from both Alphaproteobacteria and Cyanobacteria (Figs. 2 and 3b), potentially illustrating the importance of pyruvate and ribose phosphate metabolisms in these endosymbioses.

Contributions to LECA CCM from other prokaryotic lineages. Besides contributions of the asgardarchaeal host and the alphaproteobacterial endosymbiont to the LECA proteome, several phylogenetic trees indicate donations from other prokaryotic lineages, some with good support (>95% UfBoot2). In some cases, these donations included enzyme families that lacked respective homologues in Asgardarchaeota and Alphaproteobacteria. Examples comprise phylogenies of glycolytic enzymes (such as hexokinase (HK), 6-phosphofructokinase 1 (PFKA), GAPDH, PGK and PK), enzymes involved in the PPP (glucose-6-phosphate 1-dehydrogenase (G6PD), 6-phosphogluconolactonase (PGLS), RPE, TKT and transaldolase (TAL)), as well as TCA cycle enzymes (CS, IDH1 and FUMA/B) (Figs. 2 and 3c). Potential donors that we identified included Chlamydia (TAL), Planctomycetota–Verrucomicrobiota (PGK, G6PD and RPE), Fusobacteriota (PK), Cyanobacteria, Dependentiae (for two independent donations of LDH) and Chloroflexota (CS) (Fig. 3c and Supplementary Figs. 12, 13, 17, 29 and 36). However, several phylogenies displayed a mixed composition of sister groups which hindered the identification of the donor for the respective clade (for example, GAPDH; Supplementary Fig. 7). Hence, we tallied for each LECA orthogroup the occurrence of prokaryotic taxa in its sister group, which, besides a clear archaeal and proteobacterial signal, also revealed the presence of recurrent phyla in these sister groups, including Myxococota–Desulfobacterota, Bacteroidota and Acidobacteriota (Fig. 3e). These examples suggest that prokaryotes other than Asgardarchaeota and Alphaproteobacteria have contributed to the assembly of CCM during and after eukaryogenesis.

Gene families of unresolved origins. In several (at least 12) phylogenetic reconstructions, it was not possible to clearly denote LECA clades because of paraphyletic branching of eukaryotic and prokaryotic sequences resulting in unresolved sister groups. While the phylogenetic signal was limited in some cases (phosphoglycerate mutases GPMA and GPMB or 6-phosphogluconolactonase (PGL)), others recovered consistent and robust topologies across a range of datasets and analyses; that is, glucokinase (GLK), GPI, PGD, EDA/D, PRPS1 and ACS; Supplementary Figs. 1, 3, 16, 22 and 26). For example, our phylogeny of PRPS recovers an unresolved eukaryotic group (PRPS1), that might be derived from archaeal PRPS. However, this is speculative because of the presence of interspersed bacterial groups (Fig. 3b, Supplementary Fig. 22 and Supplementary Discussion). Similarly, GPI phylogeny is consistent with previous work that also resolved paraphyletic clades for eukaryotic homologues[80,81] (Supplementary Fig. 3). Our investigation of the conservation of spliceosomal introns[82] across the MSA of eukaryotic GPI showed several conserved intron positions across eukaryotic clades, suggesting that this enzyme may in fact have been present in LECA (Mailin probability >0.5; Supplementary Fig. 3c and Supplementary Discussion). On the other hand, homologous recombination events between paralogues of different origins may explain some of the observed patterns[78] (for example, potential recombinant region in GPI; Supplementary Fig. 3d,e and Supplementary Discussion). Thus, the evolutionary origins of these latter eukaryotic gene families remain unresolved.

CCM remodelling by transfer, loss, replacement and targeting

We next investigated post-LECA evolution of CCM enzymes including their correlative distribution across the eukaryotic tree and their predicted organellar localization as inferred from organelle targeting sequences. The analysis of CCM enzyme distribution across the tree revealed that orthogroup repertoires vary between distinct eukaryotic clades (Fig. 2b). We identified both cases of independent replacement and differential retainment of isofunctional enzymes (for example, HK/GLK/ADPGK, ALDO/FBA, PGMA/PGMB/GPMI/APGM, RPIA/RPIB, PDH/POR, ACS/ACDAB, ACO/ACO2, IDH1/IDH2/IDH3 and FUMAB/FUMC; Fig. 2b). The evolutionary history of enzymes of inferred asgardarchaeal origin, such as ADPGK, APGM and RPIA, suggests that these genes were present in LECA but subsequently replaced in some eukaryotic taxa by horizontally acquired homologous or analogous enzymes of bacterial origin (HK, PGMA/B and RPIB, respectively) (Fig. 2b). A correlation network analysis (±0.5 phi coefficient cut-off; Methods) of orthogroups and lifestyle characteristics (for example, anaerobic, primary and secondary endosymbiosis) suggests that CCM enzyme repertoires partially reflect eukaryotic lifestyles (Fig. 4a, Extended Data Fig. 4 and Supplementary Discussion). Correlated distributions between photosynthetic eukaryotes (by/of primary and secondary endosymbioses) are mainly related to EMP/EDP and PPP, while correlations regarding aerobic/anaerobic lifestyle usually involved pyruvate/acetate conversions and TCA cycle enzymes (Fig. 4a). Phylogenies of POR, ACDA/B, GPI, FBA and RPIB, displayed orthogroups involving anaerobic eukaryotes (Fig. 4a), suggesting adaptations to anoxygenic conditions.

Most enzymes of the EMP and PPP as well as the key enzymes of the reverse TCA enzymes, do not encode obvious targeting signals, whereas most of the enzymes involved in pyruvate conversions and the TCA appear to be targeted to mitochondria (Fig. 4b). Nevertheless, exceptions exist, indicating potential sub- or neo-functionalization of certain enzymes. For instance, PGMA and PGMB are typically found in both the cytoplasm and mitochondria/chloroplast, whereas their analogous enzymes, GPMI and APGM, do not exhibit mitochondrial targeting sequences (Fig. 4b). Likewise, in agreement with their general targeting patterns, MDH1 is generally associated with mitochondrial functions, whereas MDH2 tends to be associated with cytoplasmic activities[74], although the reverse is true in some taxa (Fig. 4b and Extended Data Fig. 3). Not all proteins of alphaproteobacterial origin are targeted to the mitochondria (for example, ALDO and TPI) and conversely some enzymes of non-alphaproteobacterial origin appear to have mitochondrial targeting signals (for example, CS, ACNB and IDH1/2; Figs. 3 and 4b), illustrating retargeting of CCM enzymes. The following two cases exemplify the complexity of post-LECA retargeting of CCM: chloroplast and mitochondrial glycolysis (Fig. 4b and Extended Data Fig. 5). In Archaeplastida, the genes coding for EMP (and PPP) enzymes are targeted to the cytoplasm and chloroplast, respectively[83]. In particular, our results highlight the frequent duplication and subsequent relocation of 'nuclear' genes to the photosynthetic organelle (Extended Data Fig. 6 and Supplementary Discussion). Similarly, the parallel glycolysis in cytoplasm and mitochondria described in SAR[84,85], appears to be specific to secondary endosymbionts involving the lower glycolysis, between TPI and PK (Extended Data Figs. 5 and 7 and Supplementary Discussion). Therefore, the distributions of targeted proteins across the CCM enzymes illustrate the general compartmentalization of these pathways in eukaryotic cells. However, the targeting of proteins is not always in agreement with their origins, suggesting an ongoing process of retargeting during the evolution of eukaryotes[86,87].

Discussion

Our phylogenetic analyses demonstrate that a complete set of eukaryotic CCM enzymes was probably present in LECA. These enzymes originated from a variety of sources, including not only contributions from the alphaproteobacterial symbiont but also from the asgardarchaeal host and other prokaryotic donor lineages (Figs. 2a, 3e and 5). We found six putative contributions from Asgardarchaeota to the CCM of LECA, within the EMP and PPP (Fig. 2a): ADPGK, GPMI, APGM, ENO, PK and RPIA, which is in contrast to previous work postulating that

https://doi.org/10.1038/s41559-025-02648-0

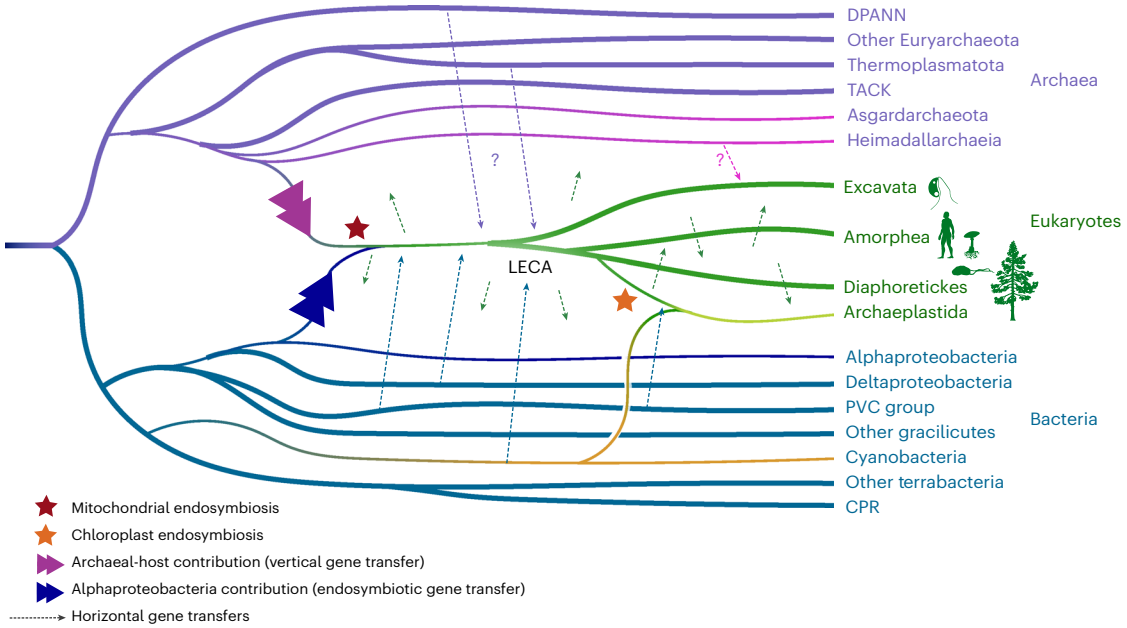


Fig. 5 | Evolutionary model for the origin of CCM during eukaryogenesis. Schematic representation of tree of life illustrating the evolutionary history for the assembly of CCM in eukaryotes. Please note that we only depict selected major routes of HGT, which is pervasive in both eukaryotic, archaeal and bacterial evolution. PVC, Planctomycetes, Verrucomicrobia and Chlamydiae[122]; CPR, Candidate Phylum Radiation[123]. Credit: Icons from PhyloPic under a Creative Commons license CC0 1.0: protist (*Andalucia godoyi*), Matus Valach; human, T. Michael Keesey; fungus, Guillaume Dera; protist (*Colponema vietnamica*), Guillaume Dera; tree, Gabriele Midolo.

Asgardarchaeota did not contribute to eukaryotic CCM[12] or that ENO was the only eukaryotic enzyme within carbon metabolism to be of archaeal origin[88,89]. Even more salient is the potential archaeal affiliation of MDH1/2 and ACLA/B/Y which are involved in the TCA and reverse TCA cycles, and which might therefore represent archaeal host contributions that became integrated into the mitochondrial TCA cycle. With the exception of ENO, these asgardarchaeal host contributions are patchily distributed in extant eukaryotes, apparently because of independent horizontal replacement events. This, combined with limited taxon sampling of prokaryotes and microbial eukaryotes in previous studies, might explain why these contributions went undetected. Our findings on asgardarchaeal contributions to the CCM strengthen the idea that eukaryotic metabolism emerged from the integration of genes from both symbiotic partners (Fig. 5), rather than being derived solely from the mitochondrial progenitor[89].

We found 17 putative alphaproteobacterial contributions, most of which are predicted to operate in the mitochondria (except for TPI, ALDO and PRPS; Fig. 2a). This finding is reminiscent of the evolutionary mosaicism previously reported for another essential process in eukaryotes, iron–sulfur cluster biosynthesis, in which the mitochondrial steps are predominantly alphaproteobacterial in origin, while the cytosolic steps are carried out by enzymes of varying evolutionary affinities[90]. While eukaryotes appear to have several CCM genes acquired from different individual bacterial taxa other than alphaproteobacteria, no additional dominant source of gene donations is apparent (Fig. 3e). We nonetheless note that a substantial number of phylogenies displayed a mixed composition of sister groups. This may be due to lack of phylogenetic signal, undersampling (that is, lack of sequence data) of relevant prokaryotic taxa and the ongoing evolution of, and HGT within and between, archaea, bacteria and eukaryotes[91–96]. Furthermore, we identified certain lineages within the sister groups that have previously been suggested to have exchanged genes with stem eukaryotes, such as Chlamydiota[45] and Myxococcota[97]. This diversity of potential donors highlights the mosaicism of the CCM in eukaryotes including contributions from additional prokaryotic sources.

The cyanobacterial contributions representing EGTs from the chloroplast to the Archaeplastida ancestor operate in the EMP and PPP (Fig. 2a), which are connected to the Calvin cycle[83]. The evolutionary origins of both chloroplast and cytoplasmic versions of the EMP and PPP in Archaeplastida show a general prevalence of nuclear gene duplications over the genes originating from the chloroplast. The predominant process appears to be one in which nuclear genes were duplicated, with one copy relocated to the photosynthetic organelle, which might have promoted the genome reduction of the endosymbiont[78,86,98]. Similarly, while the targeted localization of glycolytic enzymes to the mitochondria has previously led to the suggestion of an endosymbiotic origin of glycolysis[84,85], our work does not support this conclusion. Instead, our data indicate that CCM enzymes have been retargeted between cytosol, mitochondrion and plastid many times independently during the evolution of eukaryotes, revealing an ongoing remodelling of eukaryotic CCM.

Our results show that investigating the origin of the eukaryotic metabolism is crucial to inform our understanding of eukaryogenesis and the impact of the two primary endosymbiotic events that occurred during the origin and diversification of eukaryotes. The archaeal contributions we identify are not consistent with the view that eukaryotic metabolism is exclusively of bacterial origin[34–39]. Instead, they suggest that eukaryotic CCM is the result of an integration of host and symbiont contributions and continuous HGT (Fig. 5). The observation that most enzymes of archaeal ancestry are cytosolic and operate in the EMP and PPP, while genes of alphaproteobacterial origin function in the TCA within the mitochondrial organelle (Fig. 2a), is consistent with symbiogenetic models of eukaryogenesis: that is, models that invoke an archaeal origin of the eukaryotic cytoplasm and an alphaproteobacterial origin of the mitochondrium[15,21,24,26,28]. Specifically, our results support the view of syntrophic interactions between host and endosymbiont, in which the archaeal partner produced reducing equivalents by the degradation of organic substrates via glycolysis which, in the absence of a suitable electron acceptor, were shuttled to a bacterial symbiont which contributed a TCA cycle and an electron transport chain[24–26,30,99]. While we could not identify a third dominant

donor lineage, our results suggest that some CCM enzymes present in LECA have other phylogenetic origins among prokaryotes which may be due to transient interactions with other prokaryotes before and during eukaryogenesis. Given our results and the previously undetected asgardarchaeal host contributions to the eukaryotic CCM, we expect future studies analysing the gene origins of additional metabolic pathways to further inform symbiogenetic models for the origin of the eukaryotic cell.

Methods

Dataset construction

Initial proteome selection, annotation and redundant filtering in the core dataset. We assembled a representative and balanced dataset of selected proteomes comprising 483 archaea, 487 bacteria (5 archaeal and 95 bacterial phyla) and 224 eukaryotic proteomes, which we refer to as core database (Supplementary Data 1). We collected representatives of all major eukaryotic clades available in 2021, selected on the basis of proteome quality (that is, completeness and prevalence of contamination). For archaea and bacteria, we preferentially selected type strains and high-quality metagenome-assembled genomes (MAGs) (based on completeness (>90%) and contamination (<5%) scores). In addition, we added MAGs representing taxa that did not fulfil our stringent quality criteria such as genome-reduced DPANN and CPR which otherwise would not be present in our core database. Each proteome was annotated with eggNOG-mapper v.2.1.4-2 (MMseqs search mode[100,101]), KOFAM_SCAN v.1.3.0 (ref. 102) (-f mapper-one-line, *e* value 1×10^{-3}) and HMMSEARCH (HMMER.3.2.3 (ref. 103), *e* value $< 1 \times 10^{-3}$, selecting best *i*-value hit) against KO.hmm database[104]. We also performed DIAMOND v.2.0.6 (ref. 105) protein sequence searches against NCBI_nr release 244. To identify sequences for metabolic gene trees, we primarily used Kegg orthology (KO; Supplementary Data 2) annotations, prioritizing KOFAM classifications. In instances where KOFAM annotation was absent, we relied on HMMSEARCH annotations. The respective sequences were additionally annotated with TargetP v.2.0 (ref. 106).

Eukaryotic proteomes were downloaded and manually selected from EukProt v.3 (ref. 107). As this selection includes a variety of sequencing methods (genomes, transcriptomes and single-cell genomes), redundant and truncated sequences were filtered out uniformly. For each proteome, we first used MMseqs2 (options easy-cluster, --cluster-mode 2, --cov-mode 1, -c 1 --min-seq-id 0.95; ref. 100) and, then, used a custom script (read_clusters_mmseqs_declusterization.py) to redefine clusters.

Curation of proteomes from eukaryotic contaminations. We performed phylogenies of eukaryotic phylogenetic markers (see below) and identified prominent contaminations in the proteomes of some taxa in our dataset (Supplementary Data 3). Among others, these seem to be a result from difficulties in obtaining axenic cultures (for example, Telonemia[108]). To detect and filter out these contaminant sequences, we implemented the following workflow: first, we clustered protein families using Broccoli v.1.2.1 (ref. 109), using representative non-redundant eukaryotic proteomes. For each orthogroup, we aligned sequences with MAFFT-auto v.7.453 (ref. 110) and trimmed the MSA with trimAl 1.4.22 (ref. 111) (-gt 0.2), removing sequences with coverage <35% (custom script). We then used FastTree v.2.1.11 (ref. 112) (-lg) for inferring the phylogeny of each orthogroup. Finally, we used a custom ETE[113] script to identify contaminations defined as cases in which certain eukaryotic taxa formed a monophyletic group together with the known contaminants. Specifically, the following contaminants were removed: kinetoplastids sequence data were detected in several eukaryotic proteomes including *Lapot gusevi*, Colponemids and Telonemia, among others, and Apusomonadida sequences were detected in proteomes of *Choanocystis* sp. and *Colponema vietnamica*. For kinetoplastid contamination, truncated contaminant sequences remained after this filtering and, thus, we additionally filtered out those sequences that were taxonomically assigned to kinetoplastids given the National Center for Biotechnology Information (NCBI) and EggNog annotations (Supplementary Data 3).

KO homologies. Single KO families are not always sufficient for inferring deep evolutionary history of enzymes because they are sometimes defined on relatively shallow levels. Therefore, we inferred the homology across KO families and combined homologous families when necessary (Supplementary Data 2). We clustered all sequences from the core dataset by KO annotation and further analysed those KO families with more than ten sequences. Specifically, sequences for each KO were aligned with MAFFT-auto v.7.453 (ref. 110), trimmed using trimAl 1.4.22 (ref. 111) (-gt 0.35) and again curated with trimAl (-maxidentity 0.85 -seqoverlap 80 -resoverlap 0.5). Next, we made individual hidden Markov models (HMMs) with the HH-suite 3.1.0 package[114], using HHMAKE (-M 50). We combined all the resulting KO.hhm (14,744) into a single HH-suite database. Then, we performed HHSEARCH of KO.hhm of interest against our HH-suite database (Supplementary Data 2). Finally, we merged those KOs that were relevant for inferring the evolutionary history of certain families.

Investigating the origins of LECA clades using expanded dataset. To improve identification of prokaryotic origins of eukaryotic KO families, we searched potential LECA gene families (preliminarily identified from initial trees, see below) against a broader set of prokaryotic (NCBI-GTDB) and virus (NCBI) proteomes. We assembled a local dataset including all translated genomes from NCBI that have GTDB[49] annotation and whose genome completeness was >75% and genome contamination was <5% (a total of 187,681 prokaryotic proteomes which were over-represented in phyla such as Proteobacteria, Firmicutes and Actinobacteria among others; Supplementary Data 1). Additionally, we added viruses from NCBI (a total of 44,889 viral proteomes). We refer to this database as the expanded dataset. The workflow was as following: we first screened potential LECA clades across the preliminary phylogenies of CCM enzymes (see below) and performed respective HMM protein models using exclusively eukaryotic sequences. Then, we performed HMMSEARCHES (*e* value 1×10^{-5}) of these eukaryotic HMMs against the expanded prokaryotic and viral datasets. To avoid over-representation of taxa, for each HMMSEARCH we selected the top 15 sequences for each taxonomic class until we collected a total of 150 sequences. Then, we added these sequences to our original set of sequences from the core dataset (removing redundant sequences at 97% of identity threshold using trimAl). These extended searches provided potential donors that were overlooked in the core dataset (for example, LDH phylogeny).

Phylogenetic analyses

Eukaryotic tree of life phylogenies. The eukaryotic tree of life was reconstructed by the concatenation of the alignments of phylogenetic markers that were carefully and individually assessed and curated through iterative phylogenetic reconstructions. We first assembled protein HMMs (MAFFT-auto v.7.453 (ref. 110), trimAl 1.4.22 (ref. 111) -gt 0.4, HMMBUILD[103]) using the sequences for 320 markers provided in ref. 50. For each phylogenetic marker HMM, we performed an HMMSEARCH (*e* value 1×10^{-15}) against the eukaryotic proteomes and extracted the top ten sequences of each taxon sorted by individual *e*-value per domain. We performed an initial phylogeny using MAFFt-auto v.7.453 (ref. 110), trimming with trimAl 1.4.22 (ref. 111) (-gt 0.70) and FastTree v.2.1.11 (ref. 112) (-lg) to identify the orthogroup in question and remove spurious and/or long-branching sequences. Then, we performed two other rounds of phylogenies using MAFFT-L-INS-i v.7.453 (ref. 110), BMGE 1.12 (ref. 115) (-h 0.55), MSA cover >35% and built the gene tree with IQ-TREE 2.1.2 using ultrafast bootstrap with the best-fitting empirical or mixture model[116,117] (-bb 1000 -mset LG -madd LG + C10,LG + C20,LG + C10 + R + F,LG + C20 + R + F). These two

rounds were used to identify and remove contaminating sequences and select a single orthologue per taxon on the basis of the phylogenetic position and the sequence length relative to the total alignment length (note that three phylogenetic markers were excluded because of low phylogenetic resolution). We finally concatenated 317 markers which were individually aligned with MAFFT-L-INS-i v.7.453 (ref. 110) and trimmed with BMGE 1.12 (ref. 115) (-h 0.55). Phylogenetic analyses were based on IQ-TREE 2.1.2 (ref. 117) (see below). Taxa with a concatenation coverage <50% as well as fast-evolving taxa such as Microsporidia were excluded for analyses focusing on the eukaryotic tree of life (Fig. 1 and Supplementary Data 1).

We first reconstructed a phylogeny using corrected UBFoot2 and the LG + C60 + G mixture model (-mset LG -madd LG + C60 + G --score-diff all -bb 1000 -bnni) with IQ-TREE 2.1.2 (refs. 116,117). We then gradually removed heterogeneous sites using 'alignment_pruner.pl' script (--chi2_prune 0-0.9; https://github.com/novigit/davinciCode/blob/master/perl), followed by phylogenetic inferences using IQ-TREE 2.1.2 (refs. 116,117) (-mset LG -madd LG + C60 -bb 1000). We addition ally reduced the MSA to 148 selected eukaryotes and performed a Bayesian phylogeny using PhyloBayes 3 (ref. 118) (-catfix C60, -gtr) although the chains did not converge (11,700 generations, max_dif=1, meandif=0.03).

CCM enzyme phylogenies. We used the metabolic maps of glycolysis, PPPs, Entner–Doudoroff pathway, pyruvate metabolisms and TCA cycle provided by KEGG (https://www.genome.jp/kegg/pathway.html) and determined their distribution across our core dataset to select those KOs that were present in eukaryotes (Supplementary Data 2). Instances such as glyceraldehyde-3-phosphate ferredoxin oxidoreductase and ketoglutarate dehydrogenase/multifunctional 2-oxoglutarate metabolism enzyme among other enzymes, were not found in eukaryotes and excluded in downstream analyses.

To reconstruct refined gene tree phylogenies, we performed three main steps (Extended Data Fig. 2). In the preliminary phase, we built an initial and curated phylogeny using the core dataset, by using a strict MSA covering threshold (>80%), visual inspection of the MSA and removing terminal long branches (that is, branches longer than six times the mean of all the terminal branch lengths, as assessed using the script 'read_terminalbranchlength.py'). Final trees were obtained with IQ-TREE 2.1.2 using the best models and optimized UfBoot2 (-mset LG -madd LG + C20 + G + F -bb 1000 -bnni -alrt).

Then, we manually inspected the trees and identified potential LECA clades (including cases such as GPI and PGD) to build a eukaryotic-specific HMM. These HMM were then used for the extension phase in which we made HMMSEARCHES of each 'LECA' HMM against our local NCBI database including prokaryotes and viruses and add the 150 top sequences to our final set of sequences obtained in the previous phase (see above, expanded dataset).

The final phase consisted in two kinds of reconstructions. One was based on the strict trimming (MSA cover >80%) to get consistent sister group relationship and definition of LECA clades, while the other was based on inclusive trimming (MSA cover >20%) to include truncated sequences in the absence/presence profiles across eukaryotes. Final phylogenies are based on Mafft-L-INS-i alignments trimmed with trimAl 1.4.22 (-gt 0.7) and IQ-TREE 2.1.2 using empirical and mixture models (-mset LG | -m LG + C20 + G + F -bb 1000 -bnni -alrt -nstop 500 -pers 0.2). Trees were manually rooted as described in the Supplementary Information. We preferentially used outgroup rooting, but when this was not possible, we chose an arbitrary root to ease visualization of sister groups of interest. In addition, some phylogenies required further refinements including addition of an outgroup (GLK, HK, ADPGK, ACO, MDH, LDH, SDH and LSC), extraction and phylogeny of single Pfam domains (PFK, H6PD, PGLS, ACDAB and ACLAB/Y), subselection of sequences for refined phylogenies (TPI, PK, EDD, EDA, TAL, TKT, PDHD, POR and FUMAB/C) and conservation of introns (GPI, PGD, ACS and MDH/LDH). All trees were annotated and visualized with Interactive Tree Of Life (iToL)[119]. All alignments and raw tree files are available via Zenodo[120] (https://doi.org/10.5281/zenodo.10991068).

Domain extraction for phylogenetic reconstructions. For PFK, H6PD and PGLS phylogenies, we extracted the respective Pfam domain of interest inferred with HMMSCAN. For the case of ACLAB/Y and ACDAB, we first built the respective MSAs including all homologues (fused and separate genes) using MAFFT-L-INS-i v.7.453 and trimAl 2.1.2 (-gt 0.4). Then, we split the MSA into the respective subunits (ACLA/ACLB and ACDA/ACDB) and built an HMM with HMMBUILD. Then, we aligned the set of sequences to the HMM using HMMALIGN (--trimm) and converted the HMM output file (that is, '.sto' file) into unaligned sequences which were used for phylogenic analyses. A similar approach was used for separated mitochondrial pyruvate carrier subunits (MPC1/2). Note that phylogeny of AcdB/AclA subunit in Supplementary Fig. 27 consists of the extraction of ATP-grasp Pfam domain. Final phylogenies were conducted as described above.

Analyses for shared introns. We investigated the shared spliceosomal intron positions for GPI, PGD, ACS and MDH/LDH gene families to investigate the potential monophyly of eukaryotic sequences (see GPI section in Supplementary Discussion for further contextualization). To identify the extent of conserved spliceosomal intron positions for our genes, we searched the HMM of interest against a set of proteomes previously selected for which genome data were available[121]. Then, we made preliminary trees using MAFFT v.7.453 (default), trimAl 1.4.22 (-gt 0.7) and FastTree v.2.1.11 (-lg), from which we selected the eukaryotic orthogroups of interest to investigate shared introns. We realigned the selected sequence using MAFFT-L-INS-i v.7.453 and used imapper (https://github.com/Julian-Vosseberg/imapper) to infer the table of shared intron positions. Finally, we made a phylogeny including closely related prokaryotic sequences previously obtained, using MAFFT-L-INS-i v.7.453, trimAl 1.4.22 (-gt 0.7) and IQ-TREE 2.1.2 (-m LG + C20 + G + F -B 1000 -alrt 1000 -bnni) and mapped the intron positions with sufficient conservation. For putative single-gene families such as GPI and PGD we mapped those positions with more than four shared introns in the same phase relative to their codon, while for putative paralogous gene families such as ACS and MDH we mapped those positions that shared introns between two subfamilies and where each of them contained at least four taxa sharing the respective intron. In addition, to obtain the probabilities for the presence of introns in LECA, we used Malin[75], a maximum-likelihood analysis of intron evolution and conservation, using as input the species tree and intron gain/loss rates provided by ref. 121 and an intron table generated with imapper scripts (Supplementary Data 5).

Topology support along MSA partitions of GPI. In the MSA of GPI, we identified a shared intron between Cyptophyceae and chloroplast clade. We made phylogenies of 20 positions downstream and upstream the intron position (at position 1,915) and the rest of the carboxy terminus, providing different topologies and suggesting a recombination event between nuclear and plastid paralogues. To verify that topology is specific to the recombined region, we made a subselection of 67 representative sequences and aligned them using L-INS-i v7.453 and trimmed with BMGE 1.12 (-h 0.55). Then, we split the MSA into partitions of 14 positions, to span the potentially recombined region identified visually. We performed phylogenies of each partition using IQ-TREE 2.1.2 (-bb 1000) with LG + G + F and C20 + G + F models. Finally, we read the *ufboot files with a custom ETE4 script, to investigate the Ufboot2 of the topologies of interest: Cryptophyceae + Chlamydia or Cryptophyceae-only sequences branching monophyletically with plastid (Archaeplastida + Cyanobacteria) or with potential LECA paralogues.

Orthogroup definition and correlations

For the definition of orthogroups, we manually selected the sequences forming a monophyletic clade in the respective trees built using inclusive trimming (see above). We compared those with trees build using strict trimming to identify and add sequences that were not correctly placed in the expanded datasets, for example, Metamonada in the PDHD phylogeny. These manually selected orthogroups were used for plotting the phylogenetic profile (in Fig. 2b) as well as for plotting the presence of targeting signals (Fig. 4b). To infer the correlative distribution of these orthogroups, we converted the orthogroup distribution table into absences (0) and presences (1) and inferred phi correlation coefficient using the sklearn.metrics.matthews_corrcoef python function.

Ethics statement

This research did not involve animals or humans and no new data have been generated. Furthermore, the information provided here does not pose a threat to public health, safety or security, animals, plants or the environment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All genomic data of Archaea and Bacteria analysed are available at NCBI, while all eukaryotic proteomes were downloaded from EukProt v.3 and are provided together with the annotations via Zenodo at https://doi.org/10.5281/zenodo.10991068 (ref. 120). Data generated in this study including single-gene tree analyses, concatenated phylogenies and manual annotations (that is, sequence files, alignments and tree files, compositions of orthogroups and sister groups and so on) are also available via Zenodo[120]. Public databases are available as follows: EggNog annotations were obtained with Eggnog-mapper 2.1.4-2 (https://github.com/eggnogdb/eggnog-mapper), KOFAM annotations and KO profiles downloaded from the KEGG Automatic Annotation Server in 2021 (https://www.genome.jp/tools/kofamkoala/), the NCBI proteomes were downloaded in November 2021 (https://ftp.ncbi.nlm.nih.gov/genomes/) using taxonomic annotations from GTDB (https://data.gtdb.ecogenomic.org/) and eukaryotic proteomes were downloaded from EukProt v.3 (https://doi.org/10.6084/m9.figshare.12417881.v3). Source data are provided with this paper.

Code availability

Workflows for annotations and phylogenies and custom python scripts to analyse and parse annotation data for figure generation are available via Zenodo at https://doi.org/10.5281/zenodo.10991068 (ref. 120). We used the following published codes: https://github.com/takaram/kofam_scan/tree/master, https://github.com/novigit/davinciCode/blob/master/perl/alignment_pruner.pl and https://github.com/JulianVosseberg/imapper.

References

1. Parfrey, L. W., Lahr, D. J. G., Knoll, A. H. & Katz, L. A. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc. Natl Acad. Sci. USA* **108**, 13624–13629 (2011).
2. Eme, L., Sharpe, S. C., Brown, M. W. & Roger, A. J. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* **6**, a016139 (2014).
3. Betts, H. C. et al. Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).
4. Gueneli, N. et al. 1.1-billion-year-old porphyrins establish a marine ecosystem dominated by bacterial primary producers. *Proc. Natl Acad. Sci. USA* **115**, E6978–E6986 (2018).
5. Mahendrarajah, T. A. et al. ATP synthase evolution on a cross-braced dated tree of life. *Nat. Commun.* **14**, 7456 (2023).
6. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
7. Mills, D. B. et al. Eukaryogenesis and oxygen in Earth history. *Nat. Ecol. Evol.* **6**, 520–532 (2022).
8. Craig, J. M., Kumar, S. & Hedges, S. B. The origin of eukaryotes and rise in complexity were synchronous with the rise in oxygen. *Front. Bioinforma.* **3**, 1233281 (2023).
9. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
10. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
11. Liu, Y. et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).
12. Eme, L. et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
13. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The origin and diversification of mitochondria. *Curr. Biol.* **27**, R1177–R1192 (2017).
14. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
15. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
16. Cavalier-Smith, T. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int. J. Syst. Evol. Microbiol.* **52**, 297–354 (2002).
17. Martijn, J. & Ettema, T. J. G. From archaeon to eukaryote: the evolutionary dark ages of the eukaryotic cell. *Biochem. Soc. Trans.* **41**, 451–457 (2013).
18. Baum, D. A. & Baum, B. An inside-out origin for the eukaryotic cell. *BMC Biol.* **12**, 76 (2014).
19. Guy, L., Saw, J. H. & Ettema, T. J. G. The archaeal legacy of eukaryotes: a phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**, a016022 (2014).
20. Wang, Z. & Wu, M. Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. *PLoS ONE* **9**, e110685 (2014).
21. Koonin, E. V. Origin of eukaryotes from within Archaea, archaeal eukaryome and bursts of gene gain: eukaryogenesis just made easier? *Philos. Trans. R. Soc. B* **370**, 20140333 (2015).
22. Martin, W. F., Garg, S. & Zimorski, V. Endosymbiotic theories for eukaryote origin. *Philos. Trans. R. Soc. B* **370**, 20140330 (2015).
23. 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. B* **370**, 20140327 (2015).
24. Spang, A. et al. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol.* **4**, 1138–1148 (2019).
25. Imachi, H. et al. Isolation of an archaeon at the prokaryote–eukaryote interface. *Nature* **577**, 519–525 (2020).
26. López-García, P. & Moreira, D. The syntrophy hypothesis for the origin of eukaryotes revisited. *Nat. Microbiol.* **5**, 655–667 (2020).
27. Speijer, D. Debating eukaryogenesis—Part 1: Does eukaryogenesis presuppose symbiosis before uptake? *BioEssays* **42**, e1900157 (2020).
28. Donoghue, P. C. J. et al. Defining eukaryotes to dissect eukaryogenesis. *Curr. Biol.* **33**, R919–R929 (2023).
29. Dacks, J. B. et al. The changing view of eukaryogenesis—fossils, cells, lineages and how they all come together. *J. Cell Sci.* **129**, 3695–3703 (2016).
30. Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N. & Martin, W. F. Lokiarchaeon is hydrogen dependent. *Nat. Microbiol.* **1**, 16034 (2016).

31. Martin, W. & Koonin, E. V. Introns and the origin of nucleus–cytosol compartmentalization. *Nature* **440**, 41–45 (2006).
32. Burns, J. A., Pittis, A. A. & Kim, E. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nat. Ecol. Evol.* **2**, 697–704 (2018).
33. Baum, B. & Spang, A. On the origin of the nucleus: a hypothesis. *Microbiol. Mol. Biol. Rev.* **87**, e0018621 (2023).
34. Jain, R., Rivera, M. C. & Lake, J. A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl Acad. Sci. USA* **96**, 3801–3806 (1999).
35. McInerney, J. O., O'Connell, M. J. & Pisani, D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* **12**, 449–455 (2014).
36. Rochette, N. C., Brochier-Armanet, C. & Gouy, M. Phylogenomic test of the hypotheses for the evolutionary origin of eukaryotes. *Mol. Biol. Evol.* **31**, 832–845 (2014).
37. Pittis, A. A. & Gabaldón, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).
38. Méheust, R. et al. Formation of chimeric genes with essential functions at the origin of eukaryotes. *BMC Biol.* **16**, 30 (2018).
39. Knopp, M., Stockhorst, S., van der Giezen, M., Garg, S. G. & Gould, S. B. The asgard archaeal-unique contribution to protein families of the eukaryotic common ancestor was 0.3. *Genome Biol. Evol.* **13**, evab085 (2021).
40. Canback, B., Andersson, S. G. E. & Kurland, C. G. The global phylogeny of glycolytic enzymes. *Proc. Natl Acad. Sci. USA* **99**, 6097–6102 (2002).
41. Schnarrenberger, C. & Martin, W. Evolution of the enzymes of the citric acid cycle and the glyoxylate cycle of higher plants. A case study of endosymbiotic gene transfer. *Eur. J. Biochem.* **269**, 868–883 (2002).
42. Szklarczyk, R. & Huynen, M. A. Mosaic origin of the mitochondrial proteome. *Proteomics* **10**, 4012–4024 (2010).
43. Alsmark, C. et al. Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biol.* **14**, R19 (2013).
44. Stairs, C. W. et al. Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. *eLife* **7**, e34292 (2018).
45. Stairs, C. W. et al. Chlamydial contribution to anaerobic metabolism during eukaryotic evolution. *Sci. Adv.* **6**, eabb7258 (2020).
46. Hug, L. A. et al. A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
47. Castelle, C. J. & Banfield, J. F. Major new microbial groups expand diversity and alter our understanding of the tree of life. *Cell* **172**, 1181–1197 (2018).
48. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The new tree of eukaryotes. *Trends Ecol. Evol.* **35**, 43–55 (2020).
49. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
50. Strassert, J. F. H., Irisarri, I., Williams, T. A. & Burki, F. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat. Commun.* **12**, 1879 (2021).
51. Hampl, V. et al. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic 'supergroups'. *Proc. Natl Acad. Sci. USA* **106**, 3859–3864 (2009).
52. Rogozin, I. B., Basu, M. K., Csürös, M. & Koonin, E. V. Analysis of rare genomic changes does not support the unikont–bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Genome Biol. Evol.* **1**, 99–113 (2009).
53. He, D. et al. An alternative root for the eukaryote tree of life. *Curr. Biol.* **24**, 465–470 (2014).
54. Cerón-Romero, M. A., Fonseca, M. M., de Oliveira Martins, L., Posada, D. & Katz, L. A. Phylogenomic analyses of 2,786 genes in 158 lineages support a root of the eukaryotic tree of life between opisthokonts and all other lineages. *Genome Biol. Evol.* **14**, evac119 (2022).
55. Al Jewari, C. & Baldauf, S. L. An excavate root for the eukaryote tree of life. *Sci. Adv.* **9**, eade4973 (2023).
56. Gray, M. W. et al. The draft nuclear genome sequence and predicted mitochondrial proteome of *Andalucia godoyi*, a protist with the most gene-rich and bacteria-like mitochondrial genome. *BMC Biol.* **18**, 22 (2020).
57. Pyrih, J. et al. Vestiges of the bacterial signal recognition particle-based protein targeting in mitochondria. *Mol. Biol. Evol.* **38**, 3170–3187 (2021).
58. Galindo, L. J., Prokina, K., Torruella, G., López-García, P. & Moreira, D. Maturases and group II introns in the mitochondrial genomes of the deepest Jakobid branch. *Genome Biol. Evol.* **15**, evad058 (2023).
59. Moreira, D., Blaz, J., Kim, E. & Eme, L. A gene-rich mitochondrion with a unique ancestral protein transport system. *Curr. Biol.* **34**, 3812–3819 (2024).
60. Leger, M. M. & Gawryluk, R. M. R. Evolution: a gene-rich mitochondrial genome sheds light on the last eukaryotic common ancestor. *Curr. Biol.* **34**, R776–R779 (2024).
61. Williams, S. K. et al. Extreme mitochondrial reduction in a novel group of free-living metamonads. *Nat. Commun.* **15**, 6805 (2024).
62. Burki, F. Mitochondrial evolution: going, going, gone. *Curr. Biol.* **26**, R410–R412 (2016).
63. Bui, E. T., Bradley, P. J. & Johnson, P. J. A common evolutionary origin for mitochondria and hydrogenosomes. *Proc. Natl Acad. Sci. USA* **93**, 9651–9656 (1996).
64. Stairs, C. W., Leger, M. M. & Roger, A. J. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philos. Trans. R. Soc. Lond. B* **370**, 20140326 (2015).
65. Stairs, C. W. et al. Anaeramoebae are a divergent lineage of eukaryotes that shed light on the transition from anaerobic mitochondria to hydrogenosomes. *Curr. Biol.* **31**, 5605–5612 (2021).
66. Tikhonenkov, D. V. et al. Microbial predators form a new supergroup of eukaryotes. *Nature* **612**, 714–719 (2022).
67. Chen, X. et al. The Entner–Doudoroff pathway is an overlooked glycolytic route in cyanobacteria and plants. *Proc. Natl Acad. Sci. USA* **113**, 5441–5446 (2016).
68. Gawryluk, R. M. R., Eme, L. & Roger, A. J. Gene fusion, fission, lateral transfer, and loss: Not-so-rare events in the evolution of eukaryotic ATP citrate lyase. *Mol. Phylogenet. Evol.* **91**, 12–16 (2015).
69. Karnkowska, A. et al. A eukaryote without a mitochondrial organelle. *Curr. Biol.* **26**, 1274–1284 (2016).
70. Novák, L. V. F. et al. Genomics of preaxostyla flagellates illuminates the path towards the loss of mitochondria. *PLoS Genet.* **19**, e1011050 (2023).
71. Verhees, C. H. et al. The unique features of glycolytic pathways in Archaea. *Biochem. J* **375**, 231–246 (2003).
72. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
73. Castelle, C. J. et al. Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Curr. Biol.* **25**, 690–701 (2015).
74. Gietl, C. Malate dehydrogenase isoenzymes: cellular locations and role in the flow of metabolites between the cytoplasm and cell organelles. *Biochim. Biophys. Acta* **1100**, 217–234 (1992).

75. Csűrös, M. Malin: maximum likelihood analysis of intron evolution in eukaryotes. *Bioinformatics* **24**, 1538–1539 (2008).
76. Guy, L. & Ettema, T. J. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).
77. More, K., Klinger, C. M., Barlow, L. D. & Dacks, J. B. Evolution and natural history of membrane trafficking in eukaryotes. *Curr. Biol.* **30**, R553–R564 (2020).
78. Timmis, J. N., Ayliffe, M. A., Huang, C. Y. & Martin, W. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* **5**, 123–135 (2004).
79. Stiller, J. W. et al. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat. Commun.* **5**, 5764 (2014).
80. Stechmann, A., Baumgartner, M., Silberman, J. D. & Roger, A. J. The glycolytic pathway of *Trimastix pyriformis* is an evolutionary mosaic. *BMC Evol. Biol.* **6**, 101 (2006).
81. Grauvogel, C., Brinkmann, H. & Petersen, J. Evolution of the glucose-6-phosphate isomerase: the plasticity of primary metabolism in photosynthetic eukaryotes. *Mol. Biol. Evol.* **24**, 1611–1621 (2007).
82. Irimia, M. & Roy, S. W. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb. Perspect. Biol.* **6**, a016071 (2014).
83. Maeda, H. A. & Fernie, A. R. Evolutionary history of plant metabolism. *Annu. Rev. Plant Biol.* **72**, 185–216 (2021).
84. Río Bártulos, C. et al. Mitochondrial glycolysis in a major lineage of eukaryotes. *Genome Biol. Evol.* **10**, 2310–2325 (2018).
85. Liaud, M. F., Lichtlé, C., Apt, K., Martin, W. & Cerff, R. Compartment-specific isoforms of TPI and GAPDH are imported into diatom mitochondria as a fusion protein: evidence in favor of a mitochondrial origin of the eukaryotic glycolytic pathway. *Mol. Biol. Evol.* **17**, 213–223 (2000).
86. Martin, W. Evolutionary origins of metabolic compartmentalization in eukaryotes. *Philos. Trans. R. Soc. B* **365**, 847–855 (2010).
87. A. von der Dunk, S. H. & Snel, B. Recurrent sequence evolution after independent gene duplication. *BMC Evol. Biol.* **20**, 98 (2020).
88. Hannaert, V. et al. Enolase from *Trypanosoma brucei*, from the amitochondriate protist *Mastigamoeba balamuthi*, and from the chloroplast and cytosol of *Euglena gracilis*: pieces in the evolutionary puzzle of the eukaryotic glycolytic pathway. *Mol. Biol. Evol.* **17**, 989–1000 (2000).
89. Martin, W. & Russell, M. J. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos. Trans. R. Soc. Lond. B* **358**, 59–83 (2003).
90. Freibert, S.-A. et al. Evolutionary conservation and in vitro reconstitution of microsporidian iron–sulfur cluster biosynthesis. *Nat. Commun.* **8**, 13932 (2017).
91. Archibald, J. M. Gene transfer in complex cells. *Nature* **524**, 423–424 (2015).
92. Ku, C. et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* **524**, 427–432 (2015).
93. Leger, M. M., Eme, L., Stairs, C. W. & Roger, A. J. Demystifying eukaryote lateral gene transfer (response to Martin 2017 10.1002/bies.201700115). *BioEssays* **40**, 1700242 (2018).
94. Wu, F. et al. Unique mobile elements and scalable gene flow at the prokaryote–eukaryote boundary revealed by circularized Asgard Archaea genomes. *Nat. Microbiol.* **7**, 200–212 (2022).
95. Filée, J. et al. Bacterial origins of thymidylate metabolism in Asgard Archaea and Eukarya. *Nat. Commun.* **14**, 838 (2023).
96. Keeling, P. J. Horizontal gene transfer in eukaryotes: aligning theory with data. *Nat. Rev. Genet.* https://doi.org/10.1038/s41576-023-00688-5 (2024).
97. Santana-Molina, C., Rivas-Marin, E., Rojas, A. M. & Devos, D. P. Origin and evolution of polycyclic triterpene synthesis. *Mol. Biol. Evol.* **37**, 1925–1941 (2020).
98. Coale, T. H. et al. Nitrogen-fixing organelle in a marine alga. *Science* **384**, 217–222 (2024).
99. Speijer, D. Alternating terminal electron-acceptors at the basis of symbiogenesis: how oxygen ignited eukaryotic evolution. *BioEssays* **39**, 1600174 (2017).
100. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
101. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
102. Aramaki, T. et al. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
103. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).
104. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
105. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
106. Almagro Armenteros, J. J. et al. Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* **2**, e201900429 (2019).
107. Richter, D. J. et al. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Commun. J.* **2**, e56 (2022).
108. Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V. & Burki, F. New phylogenomic analysis of the enigmatic phylum Telonemia further resolves the eukaryote tree of life. *Mol. Biol. Evol.* **36**, 757–765 (2019).
109. Derelle, R., Philippe, H. & Colbourne, J. K. Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol. Biol. Evol.* **37**, 3389–3396 (2020).
110. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
111. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinforma. Oxf. Engl.* **25**, 1972–1973 (2009).
112. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
113. Huerta-Cepas, J., Dopazo, J. & Gabaldón, T. ETE: a python environment for tree exploration. *BMC Bioinf.* **11**, 24 (2010).
114. Steinegger, M. et al. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.* **20**, 473 (2019).
115. Criscuolo, A. & Gribaldo, S. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
116. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
117. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
118. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288 (2009).

119. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
120. Santana-Molina, C., Tom A. W., Snel, B. & Spang, A. Chimaeric origin and dynamic evolution central carbon metabolism in eukaryotes. *Zenodo* https://doi.org/10.5281/zenodo.10991068 (2024).
121. Vosseberg, J., Schinkel, M., Gremmen, S. & Snel, B. The spread of the first introns in proto-eukaryotic paralogs. *Commun. Biol.* **5**, 476 (2022).
122. Wagner, M. & Horn, M. The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance. *Curr. Opin. Biotechnol.* **17**, 241–249 (2006).
123. Brown, C. T. et al. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).

Acknowledgements

A.S. and B.S. have received support from an initiative of Utrecht University (UU) to foster collaborations between UU and NIOZ (NZ4543.11: 'The origin and diversification of eukaryotic metabolisms' to A.S. and B.S.) and thank additional collaboration partners of this project, E. J. Javaux, P. Mason and R. Hennekam. A.S. has received funding from the European Research Council under the European Union Horizon 2020 research and innovation programme (grant agreement no. 947317, ASymbEL to A.S.), the Moore–Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation 735929LPI (https://doi.org/10.46714/735929LPI to A.S. and co-principal investigators). T.A.W. and A.S. have received funding from the Gordon and Betty Moore Foundation (grant no. GBMF9741 to T.A.W., A.S. and co-principal investigators). We want to thank C. Stairs, N. Dombrowski, M. Raas, S. Tzavellas and D. Tamarit for helpful discussions on eukaryotic and archaeal/bacterial taxon selection, redundancy filtering, intron analysis and phylogenetic analyses, respectively.

Author contributions

A.S. and B.S. conceived the study. C.S.-M. performed analyses, wrote code and made figures. C.S.-M., A.S. and B.S. analysed and interpreted data. T.A.W. contributed expertise and helped to interpret results. C.S.-M. with B.S. and A.S. wrote the manuscript and supplementary materials and all authors contributed to the final version of the submitted manuscript.

Competing interests

The authors declare no competing interests.

Additional information

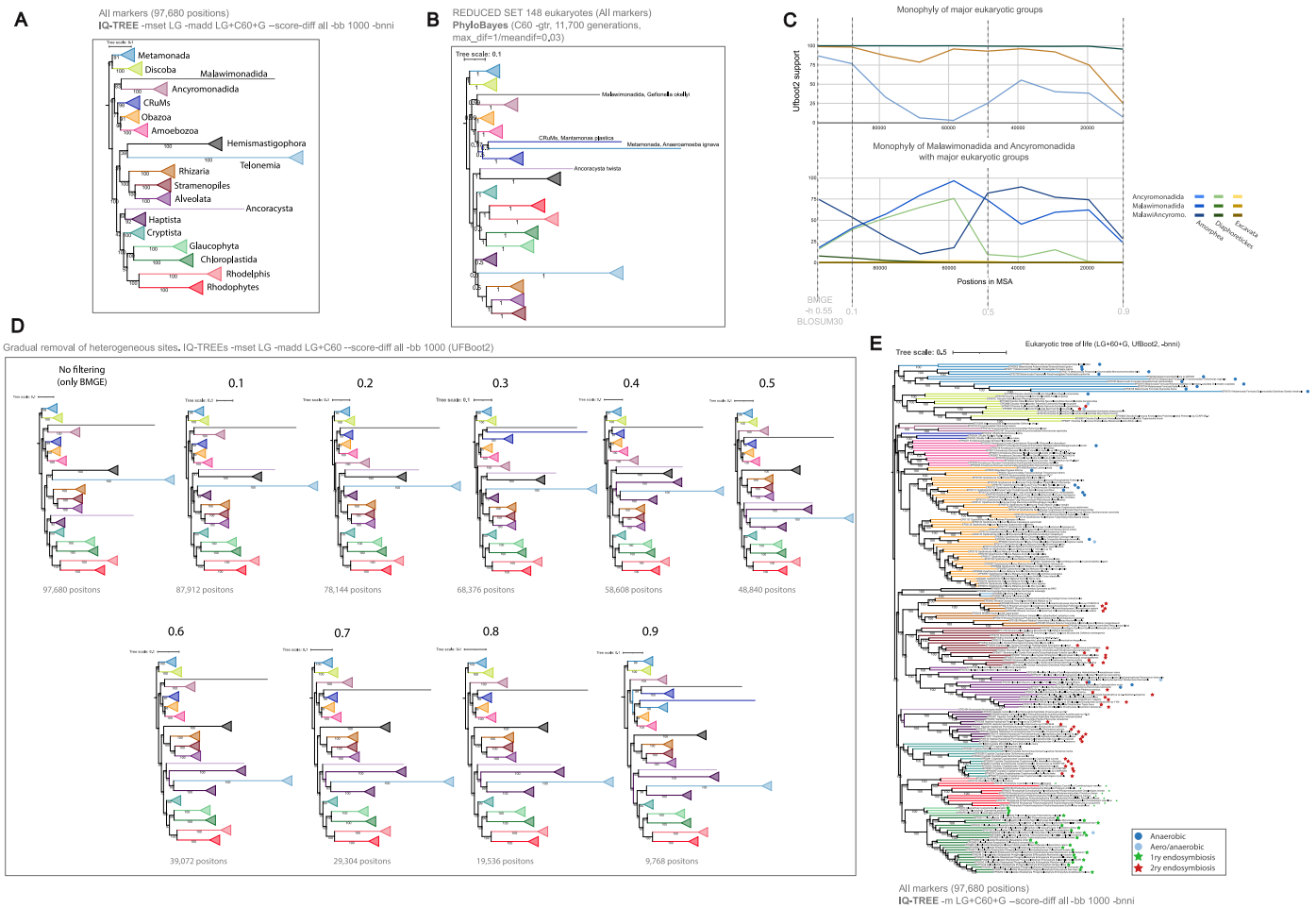
Extended data is available for this paper at https://doi.org/10.1038/s41559-025-02648-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-025-02648-0.

Correspondence and requests for materials should be addressed to Berend Snel or Anja Spang.

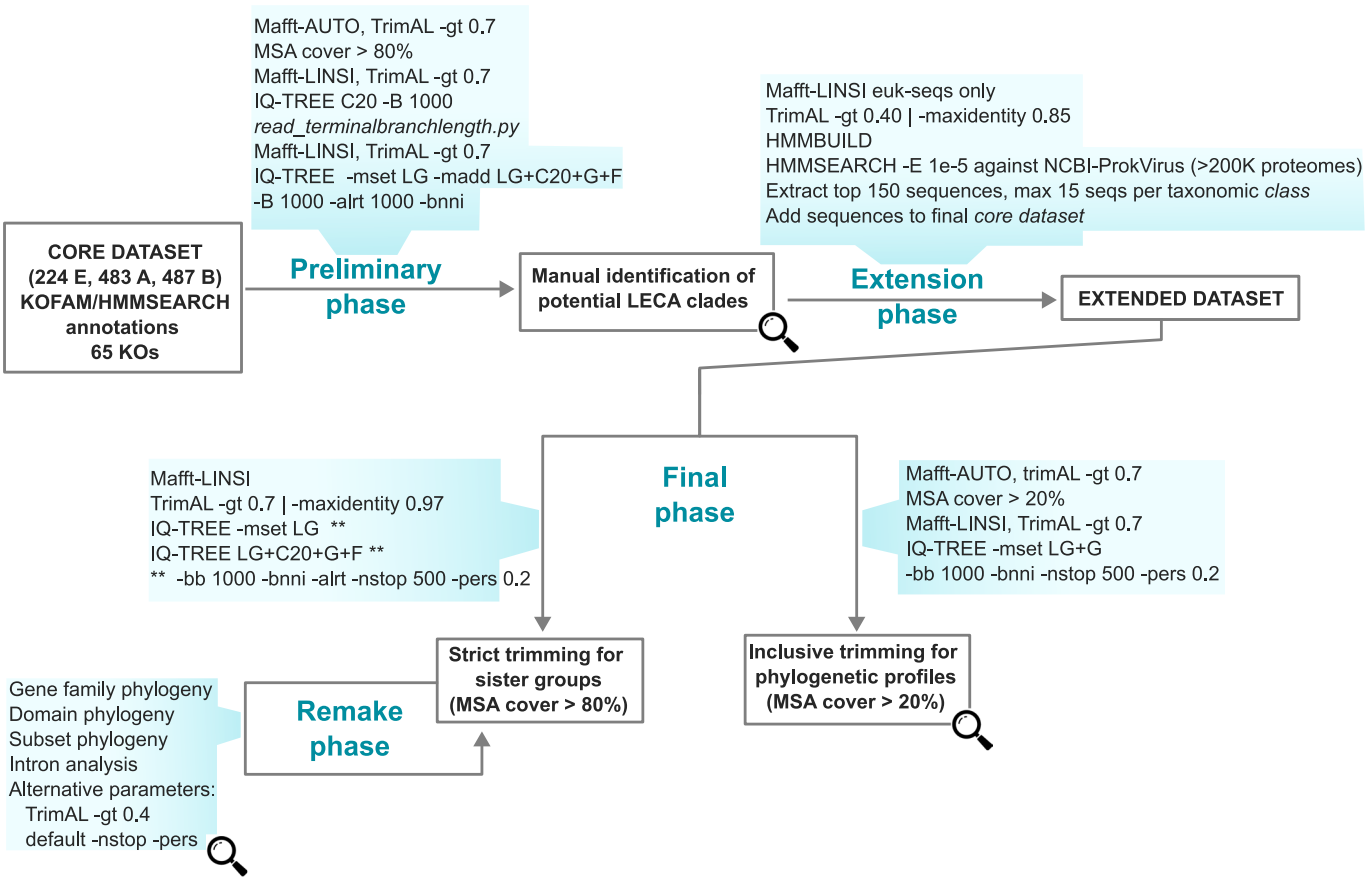
Peer review information *Nature Ecology & Evolution* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41559-025-02648-0



Extended Data Fig. 1 | Extended phylogenetic analyses for the reconstruction of the eukaryotic tree of life based on the concatenation of 317 phylogenetic markers. A) Maximum-likelihood reconstruction using IQ-TREE and ultrafast bootstraps (Ufboot2). **B)** Bayesian reconstruction of eToL using PhyloBayes. **C)** Chart depicting Ufboot2 of major groups (Excavata, Amorphea and Diaphoretickes, upper panel) and unstable groups (Ancyromonadida and Malawimonadida, lower panel) on intervals along removal of heterogeneous sites (see panel **D**). Note that Amorphea in the upper panel only includes Obozoa, Amoebozoa and CRuMs. **D)** Maximum-likelihood phylogenies of eToL after applying gradual filtering of heterogeneous sites to the concatenation (from 0.1 to 0.9). These phylogenies are used for depicting Ufboot2 values in figure S1C. **E)** Expanded view of maximum-likelihood reconstruction using IQ-TREE and corrected Ufboot2 used in Fig. 1. Phylogenetic methods and length of the concatenations are indicated in the respective panels. BUSCO values and percentage of phylogenetic markers found in each eukaryotic proteome provided in Supplementary Data 1.

https://doi.org/10.1038/s41559-025-02648-0

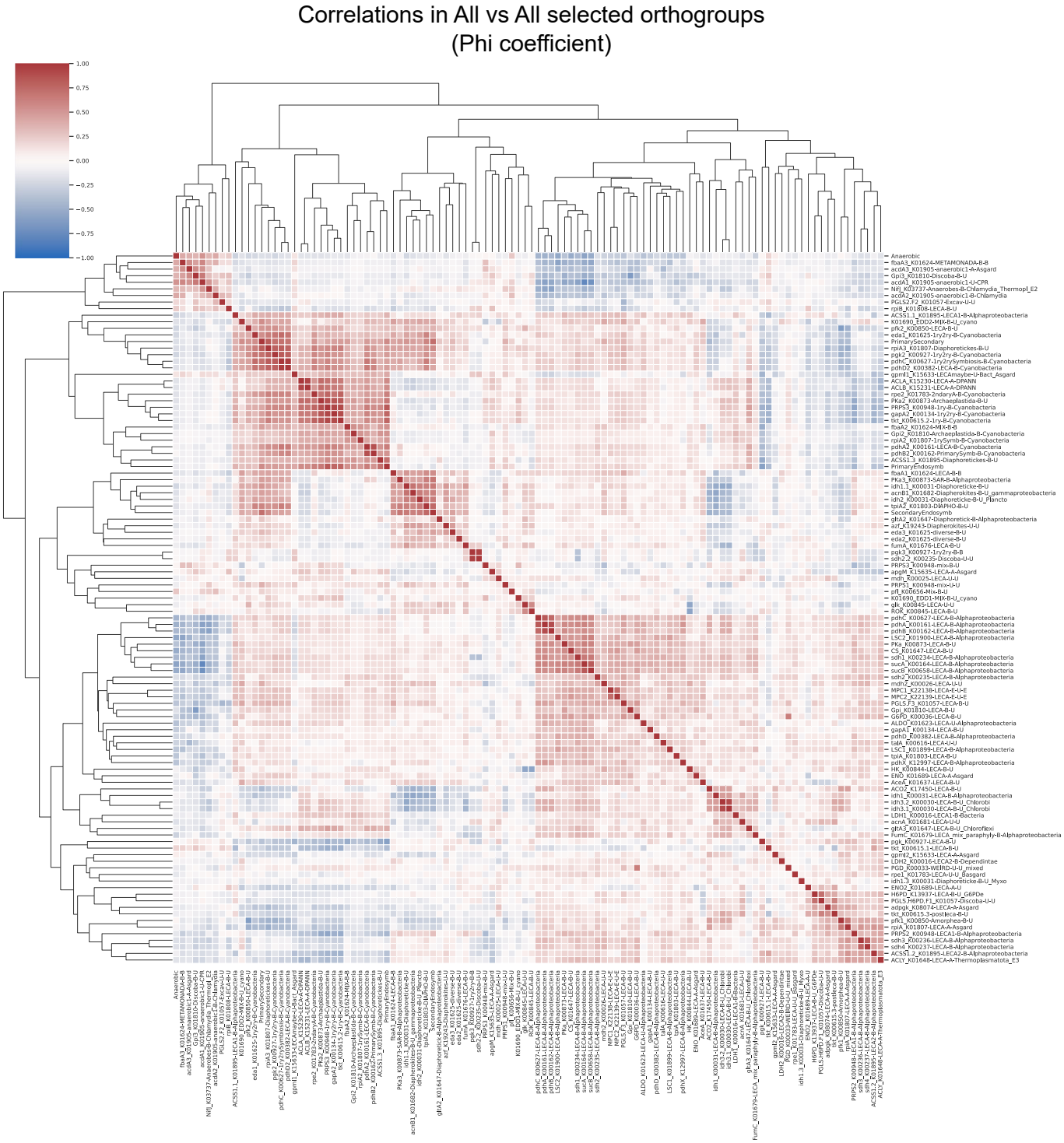


Extended Data Fig. 2 | Workflow for generating final phylogenies of central carbon metabolism enzymes for the identification of sister groups and the taxonomic composition of orthogroups. Lens indicates steps that were manually supervised. See methods for further explanation.

https://doi.org/10.1038/s41559-025-02648-0

Extended Data Fig. 3 | Extended view of phylogenetic profile for the proposed origins of selected orthogroups mapped onto the eukaryotic tree of life. Bold labels of columns indicate those that are proposed to be present in LECA. Different shapes of cells indicate presence of one or diverse sequences with the respective targeting signal obtained with TargetP (*see legend*). This is an extended view of Fig. 2b in the main text, see figure caption for additional information.

https://doi.org/10.1038/s41559-025-02648-0

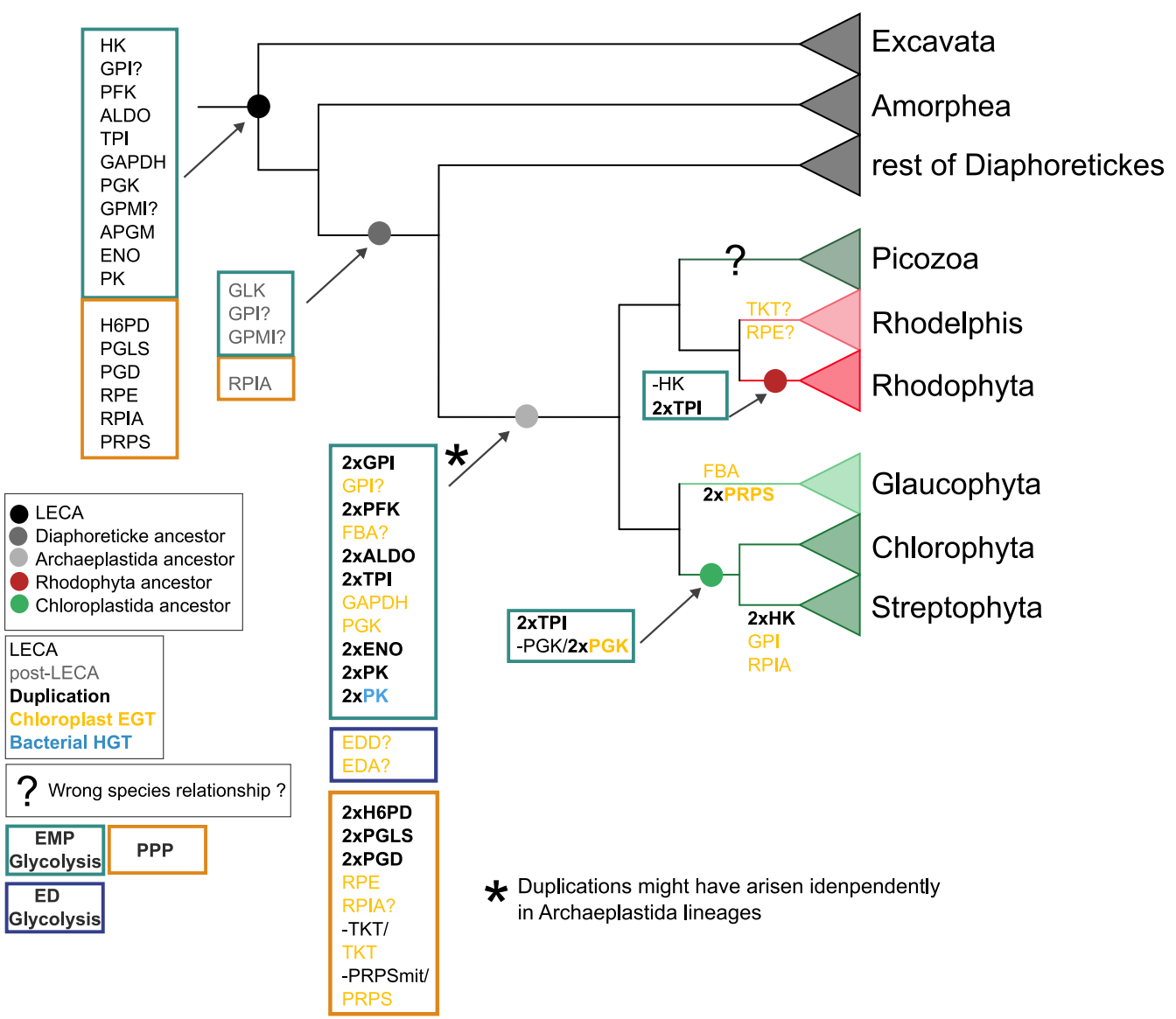


Extended Data Fig. 4 | Clustermap representing the correlations of the orthogroup distribution of CCM enzymes. The correlation was inferred using the phi coefficient (*matthews_corrcoef* function in python) and converting the matrix to 0 (absence) and 1 (presence) values. Note that the distribution of characteristic traits were included in the analysis (Anaerobic, Primary and Secondary endosymbionts). Red and blue cells indicate correlated and anticorrelated distributions respectively.

https://doi.org/10.1038/s41559-025-02648-0

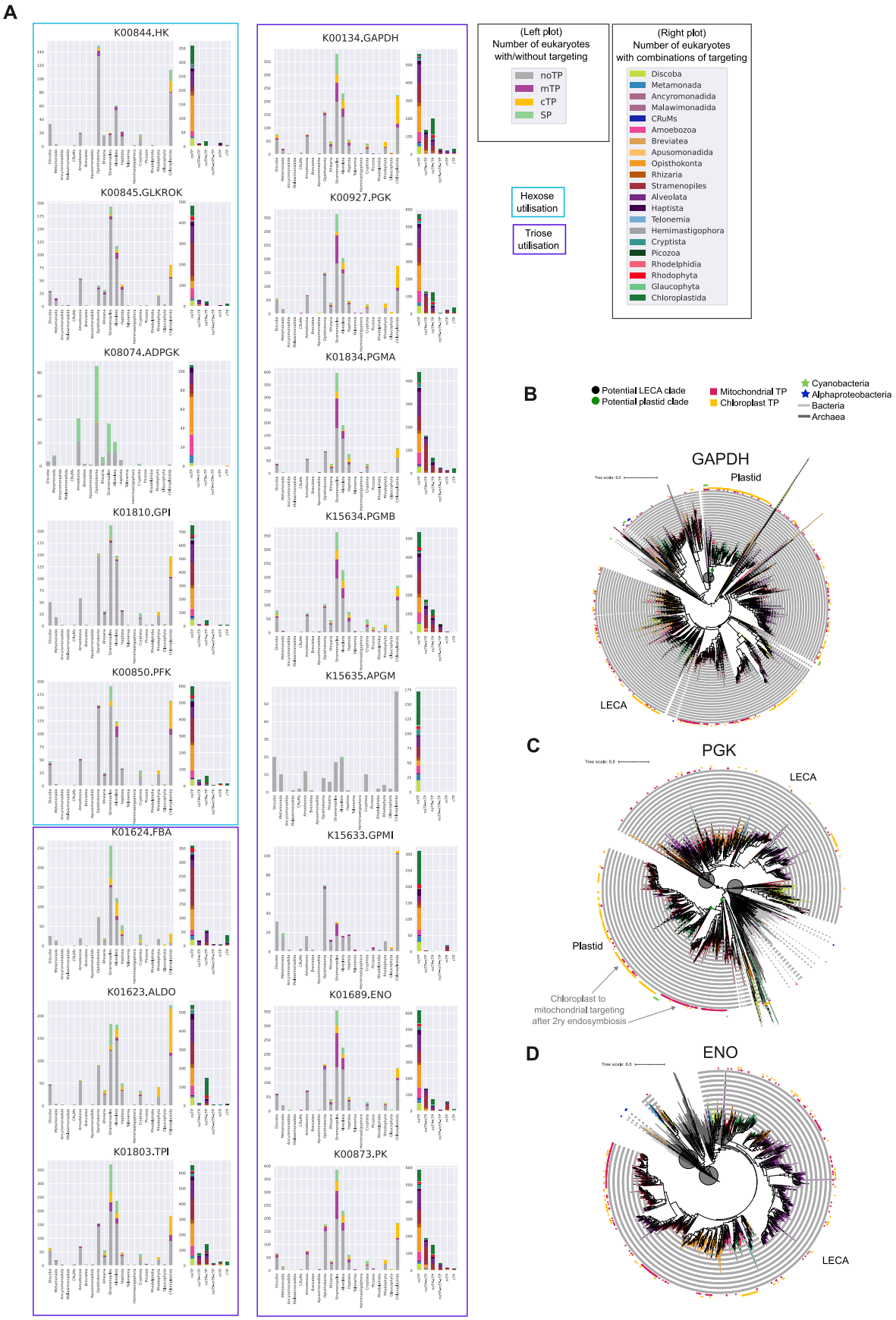
Extended Data Fig. 5 | Distribution of parallel glycolysis in the cytoplasm and in the chloroplast or mitochondria using all proteomes of EukProt v3. **A)** Parallel glycolisis targeted to the cytoplasm and chloroplast. **B)** Parallel glycolysis targeted to the cytoplasm and mitochondria. Colored cells indicate organisms that have one sequence with no targeting signal (cytoplasm) and another with the respective organelle targeting. Only those organisms with more than 3 enzymes with parallel targeting are shown.

https://doi.org/10.1038/s41559-025-02648-0



Extended Data Fig. 6 | Origins and evolution of cytoplasmic and chloroplast glycolysis and PPP mapped onto an schematic representation of the eukaryotic tree of life (zoomed in the Archaeplastida clade). The boxes encompass enzymes predicted to be present in the corresponding ancestral nodes. Tree reconciliations focus on major events, representing an approximation inferred manually. For instance, duplications within the Archaeplastida ancestor might represent independent duplications in distinct lineages. The prefix symbol minus (-) denotes potential gene loss.

https://doi.org/10.1038/s41559-025-02648-0



Extended Data Fig. 7 | See next page for caption.

https://doi.org/10.1038/s41559-025-02648-0

Extended Data Fig. 7 | Distribution of targeted glycolytic enzymes using all proteomes of EukProt v3. A) Quantification of targeted enzymes by eukaryotic group. Left panel shows the number of sequences with/without targeting signal and the right panel shows the number of eukaryotic sequences with a combination of targeting. Plots within the blue box represent those enzymes using substrates with more than three carbons while the purple box include those enzymes using substrates of three carbons (triose). Phylogeny of **B)** GAPDH, **C)** PGK, and **D)** ENO mapping the respective distribution of mitochondrial and chloroplast targeting. Branches are colored according to the eukaryotic groups, and peripheral prokaryotic groups are collapsed.

nature portfolio

Corresponding author(s): Berend Snel, Anja Spang

Last updated by author(s): 03/12/2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data of study at ZENODO, https://zenodo.org/records/10991068 EukProt v3, https://evocellbio.com/eukprot/ GTDB, https://gtdb.ecogenomic.org/downloads NCBI, https://ftp.ncbi.nlm.nih.gov/genomes/ KEGG, https://www.genome.jp/kegg/pathway.html KOFAM, https://www.genome.jp/ftp/db/kofam/
Data analysis	Custom scripts, https://zenodo.org/records/10991068 MALIN, https://www-labs.iro.umontreal.ca/~csuros/introns/malin/ eggNOG-mapper v.2.1.4-2, https://github.com/eggnogdb/eggnog-mapper MMSEQS2, https://github.com/soedinglab/MMseqs2 KOFAM_SCAN v.1.3.0, https://github.com/takaram/kofam_scan HMMER 3.2.3, http://hmmer.org/download.html HHSUITE 3.1.0, https://github.com/soedinglab/hh-suite MAFFT v7.453, https://mafft.cbrc.jp/alignment/software/ trimAL 1.4.22, https://vicfero.github.io/trimal/ BMGE 1.121, https://bioweb.pasteur.fr/packages/pack@BMGE@1.12 alignment_pruner.pl script, https://github.com/novigit/davinciCode/blob/master/perl FastTree v2.1.11, http://www.microbesonline.org/fasttree/ IQ-TREE 2.1.2, http://www.iqtree.org/#download PhyloBayes 3, https://github.com/bayesiancook/phylobayes

Broccoli v1.2.1, https://github.com/rderelle/Broccoli
imapper, https://github.com/JulianVosseberg/imapper

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All genomic data of Archaea and Bacteria analyzed are available at NCBI, while all eukaryotic proteomes were downloaded from EukProt V3, and are provided together with the annotations in our data repository at Zenodo (10.5281/zenodo.10991067). Data generated in this study including single gene tree analyses, concatenated phylogenies and manual annotations (i.e.sequence files, alignments, and treefiles, compositions of orthogroups and sister groups, etc) have also been deposited in our data repository at Zenodo (10.5281/zenodo.10991067). Public databases are available as follows: EggNog annotations were obtained with Eggnog-mapper 2.1.4-2 (https://github.com/eggnogdb/eggnog-mapper), KOFAM annotations and KO profiles downloaded from the KEGG Automatic Annotation Server in 2021 (https://www.genome.jp/tools/kofamkoala/), the NCBI proteomes were downloaded in November 2021 (https://ftp.ncbi.nlm.nih.gov/genomes/) using taxonomic annotations from GTDB (https://data.gtdb.ecogenomic.org/) and eukaryotic proteomes were downloaded from EukProt V3 (https://doi.org/10.6084/m9.figshare.12417881.v3).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	n/a
Reporting on race, ethnicity, or other socially relevant groupings	n/a
Population characteristics	n/a
Recruitment	n/a
Ethics oversight	n/a

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	In this study, we investigate the origin and evolution of central carbon metabolism of eukaryotes. We first investigated the phylogeny of the eukaryotic Tree of Life, followed by the evolution of the enzymes involved in the pathways of interest: glycolysis, pentose-phosphate pathway, pyruvate and acetate metabolism towards Acetyl-CoA and the Tricarboxylic Acid cycle. We performed extensive phylogenetic analyses using a representative set of the Bacteria, Archaea and Eukaryotes, and integrated it with additional information like protein targeting, protein domain evolution, and conservation of introns. To avoid false positive annotations due to prokaryotic contamination, we defined the orthogroups of interest by visual inspection of the final phylogenetic trees. We discuss the potential origins and distribution of each enzyme involved in the central carbon metabolism of eukaryotes. We also analyze the correlative distribution of the respective enzymes across the eukaryotes showing the different dynamics of these metabolisms in eukaryotes.
Research sample	We assembled a representative and balanced dataset of selected proteomes comprising 483 archaea, 487 bacteria (5 archaeal and 95 bacterial 353 phyla) and 224 eukaryotic proteomes, that we refer to as core database in the manuscript. In addition, we add prokaryotic sequences by doing protein sequence searches of eukaryotic orthogroups of interest against an extended dataset of 187,681 prokaryotic and 44,889 viral proteomes.

Sampling strategy	For eukaryotes we sampled proteomes from all highest taxonomic levels, and selected those more representative and with better quality. For archaea and bacteria, we preferentially selected type-strains, while high quality metagenomes assembled genomes were selected based on completeness and contamination levels. For adding prokaryotic sequences from the extended dataset, we selected the top 150 sequences of each orthogroup searches, collecting the a maximum of 15 sequences per taxonomic class.
Data collection	Prokaryotic and viral proteomes were collected from NCBI and eukaryotic proteomes were collected from EukProt.
Timing and spatial scale	We have selected all proteomes at the beginning of this project (2021).
Data exclusions	Our analyses were performed using a selected set of proteomes that combined sources from NCBI and EukProt. We have tried to select the highest quality genomes from a taxonomically representative set reflecting the known phylogenetic diversity of Archaea, Bacteria and Eukaryotes. Our strategy excluded genomes with lower quality and derived from closely related taxa.
Reproducibility	We have taken care to provide a detailed method section, supplementary information as well as data repository, which provide access to our data and code and ensure reproducibility.
Randomization	Genomic data was selected based on highest quality and widest taxonomic distance/range.
Blinding	Our paper reports phylogenetic analyses. Data were selected based on best practices, for example ensuring a representative taxon sampling. These are analyses of observational data and blinding is not part of the standard analysis protocol.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a