Diverse ancestries reveal complex symbiotic interactions during eukaryogenesis

Moisès Bernabeu[1,2,+], Saioa Manzano-Morales[1,2,+], Marina Marcet-Houben[1,2,3,+], and Toni Gabaldón[1,2,3,4*]

1. Barcelona Supercomputing Centre (BSC-CNS). Plaça Eusebi Güell, 1-3 08034 Barcelona, Spain.
2. Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain
3. CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain
4. Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

[+] These authors contributed equally and their names are listed in alphabetical order
[*] Author for correspondence: toni.gabaldon@bsc.es

Summary paragraph

The origin of eukaryotes remains a central enigma in biology[1]. Ongoing debates agree on the pivotal role of a symbiosis between an alpha-proteobacterium and an Asgard archaeon[2]. However, the involvement of other bacterial partners as significant donors of genetic material[3–5], or the role of interactions with viruses[6–8] remain contentious. To address these questions, we employed advanced phylogenomic approaches and comprehensive datasets spanning the known diversity of cellular life and viruses. Our analysis provided an updated reconstruction of the last eukaryotic common ancestor (LECA) proteome, in which we traced the phylogenetic origin of each protein family. We found compelling evidence for multiple waves of horizontal gene transfer from diverse bacterial and viral donors, with some likely preceding the mitochondrial endosymbiosis. We inferred the proteomes of the potential major donors and their functional contributions to LECA. Our findings underscore the contribution of horizontal gene transfers in shaping the proteomes of pre-LECA ancestors and hint to a facilitating role of Nucleocytoviricota viruses. Altogether, our results suggest ancient eukaryotes originated within complex microbial ecosystems through a succession of symbiotic interactions that left a footprint of horizontally transferred genes.

Main text

The sharp divide in terms of cellular complexity between eukaryotes and prokaryotes has been deemed "the greatest single evolutionary discontinuity to be found in the present-day world"[9], and the possibility of genome mergers through prokaryotic (endo)symbiosis has long been considered a stepping stone in this process[10]. The current consensus on eukaryogenesis revolves around scenarios that always involve a genome merger between an Alphaproteobacterial endosymbiont and a host with an Asgard archaeal (formally, phylum Promethearchaeota[11]) component[2]. Phylogenomics has shed light on the phylogenetic placement and potential traits of these two established partners[12–16]. However, some proposed eukaryogenesis models involve at least one additional partner[4], or even serial interactions with various non-Alphaproteobacterial symbionts acting as gene donors[3,5]. The large dominance of bacterial over archaeal contributions in reconstructed LECA proteomes[3,17], and the observation that only a small fraction of bacterial-derived proteins can be confidently traced back to Alphaproteobacteria, may hint to additional bacterial contributions[3,18]. However, the difficulty of reconstructing ancient phylogenetic relationships and the footprints of earlier horizontal gene transfer (HGT) carried by the Asgard archeal or Alphaproteobacterial partners are alternative explanations to the diversity of phylogenetic signals in LECA.

Here we asked the objective question of whether protein families within LECA could trace back to ancestors different from Alphaproteobacteria or Asgard archaea with a level of support similar to those tracing to these broadly-accepted partners. To alleviate potential artefacts arising from phylogenetic reconstruction, including the effects of unsampled lineages, contamination, and recent HGT, we used state-of-the-art methodologies and compiled curated datasets including representative proteomes with the highest possible qualities, which we purged from low-quality sequences, recent paralogues and sparsely

distributed proteins. Our results identify at least three significant signals of bacterial ancestry different from Alphaproteobacteria, and a large contribution from viruses of the Nucleocytoviricota phylum.

Reconstruction of the ancestral eukaryotic proteome

To leverage the recent explosion of genome data across the tree of life, particularly among eukaryotes, we re-inferred the LECA proteome. To minimise known methodological issues in homology and phylogenetic inference, we subsampled existing data to obtain a balanced representation across the eukaryotic tree of life (eTOL), while ensuring a tractable size and the highest possible quality (**see Methods, Supplementary Methods, Supplementary Tables 1 and 2**). We additionally curated the selected proteomes to remove low-quality and low-complexity proteins, and, given our focus on deep evolutionary nodes, we kept a single representative of clusters of recent eukaryotic in-paralogs. We replicated this procedure to generate three alternative 100-proteome datasets (eTOLDB-A, -B, and -C) that overlap in about 46% of their proteins and allow assessing the data-dependency of our results. We clustered proteins in these datasets into orthologous groups (OG), and selected those OGs containing at least five different species, and at least three out of nine eukaryotic supergroups and the two main eukaryotic stems (**see Methods and Supplementary Figures 1 and 2**) as putative descendants of LECA (LECA-OG), which were highly consistent across datasets (>96%). To further refine these families we used alignment profile searches against a broad database (broadDB) comprising order-level prokaryotic pangenomes reconstructed from over 65,000 genomes available at GTDB[19] and sequence representatives of over 1.3 million clusters of viral sequences[20] (**see Methods**). This approach ensured maximal coverage of extant diversity while minimising the impact of database biases and recent HGT. We next reconstructed Maximum Likelihood (ML) phylogenies from the LECA-OG expanded with the closest broadDB hits until reaching a maximum of 500 homologous sequences. This served to assess the monophyly of the eukaryotic proteins, which were split into different, monophyletic LECA-OGs (mLECA-OGs) if necessary. We repeated the same procedure with this new set of mLECA-OGs by building new HMM profiles, and repeating the broadDB search and phylogenetic reconstructions (**see Methods**). This resulted in a final set of mLECA-OGs (with 63% consistency across datasets), and their phylogenies in the context of their closest non-eukaryotic homologs. We functionally annotated these OGs and all proteomes used in subsequent analyses in the common framework of KEGG orthologs (KOs).

Our reconstructed LECA proteomes comprise an average of 9,472 or 5,572 mLECA-OGs (OGs hereafter) depending on the relaxed (3 supergroups) or strict (5 supergroups) criterion (**Supplementary Table 3**). We considered a relaxed consensus of 7,563 KOs consistently predicted by two or more datasets, or by one dataset and the strict criterion, and reconstructed the functional potential of LECA (**see Methods**). Our reconstruction confirms earlier studies depicting a complex LECA, covering the main features of extant eukaryotes (**Fig. 1a and b, Supplementary Discussion**). LECA had the core metabolism and machinery for nucleic acid and protein processing, endocytosis, and the processing of extracellular particles, and

contained phagosomes, lysosomes and peroxisomes. We infer a cytoplasm similar to that of extant eukaryotes with a cargo-based transport using the motor proteins (dynactin, kinesin and myosin). Conversely, the cell cycle appears rather primitive, with the presence of the basic machinery and a very limited regulation. Regarding metabolism, the low completeness of the autotrophic Wood–Ljungdahl and the Arnon-Buchanan cycle suggest a heterotrophic lifestyle. Moreover, the reconstruction lacks enzymes related to anaerobiosis (Pyruvate:ferredoxin oxidoreductase –PFO–, [FeFe] hydrogenase and enzymes involved in the rhodoquinone synthesis), suggesting that LECA was most likely aerobe (**Supplementary Tables 4 and 5**, **see Supplementary Discussion**).

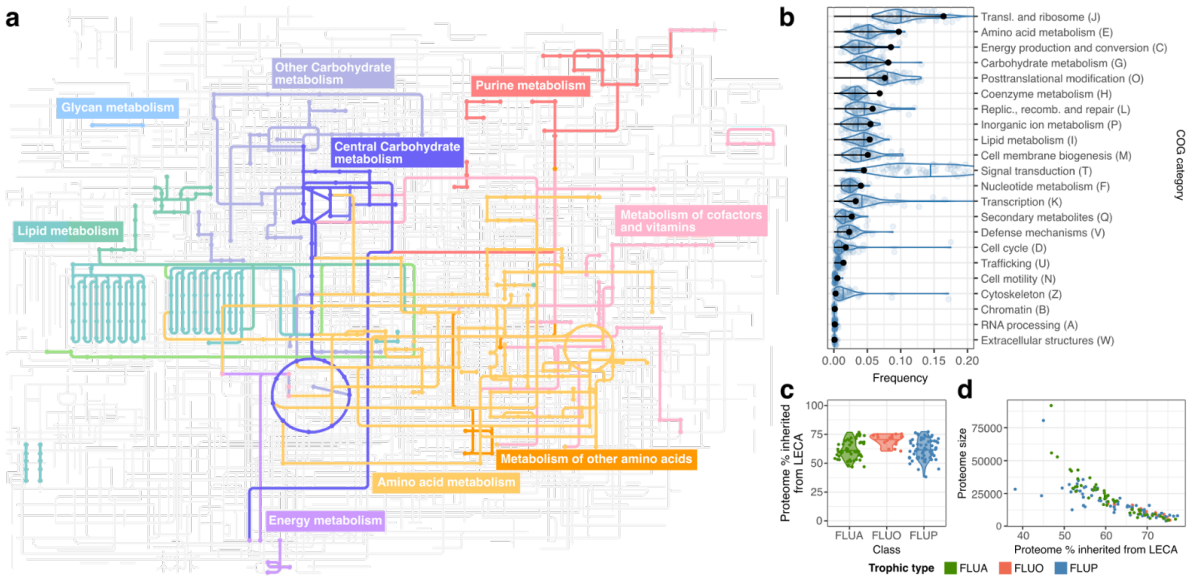

Figure 1. Reconstruction of the ancestral proteome of LECA and its metabolic features. *a) Metabolic reconstruction of the consensus LECA proteome. In light grey: the metabolic map of KEGG. In colours: the modules inferred to have been present in LECA using the consensus proteome, which accounts for the KOs present in at least 2 LECA reconstructions of different eTOLDBs or supported by a LECA group with five eukaryotic supergroups. Each colour refers to a metabolic category according to the coloured boxes. b) Frequency of COG database functional categories in the consensus LECA proteome (black) and in the 63 FLUPs (distribution in blue). c) Percentage of KEGG orthologs (KOs) shared between the ancestral proteome of LECA and extant free-living unicellular autotrophs, osmotrophs and phagotrophs (FLUAs, FLUOs and FLUPs, respectively). d) Relationship between the percentage of the proteome inherited from LECA and the extant genome size.*

We compared the reconstructed proteome with the proteomes of representative extant free-living unicellular phagotrophs (FLUPs), osmotrophs (FLUOs), and autotrophs (FLUAs) sampled across the eTOL (**Supplementary Table 6**). Although FLUPs share the inferred trophic lifestyle of LECA[2], we did not observe clear differences in terms of the absolute number of LECA KOs (**Fig. 1c**) and COG categories inherited by each trophic group (**Extended Data Fig. 1, Supplementary Table 7**). However, due to their generally smaller genomes, FLUOs tend to share a larger fraction of their proteomes with LECA (**Fig. 1d**). As this lifestyle is secondarily derived in different eukaryotic lineages, this observation suggests that FLUPs and FLUAs have expanded their proteomes with post-LECA acquisitions,

whereas subsequent proteome streamlining in FLUOs preferentially retained ancestral functions. Amid a wide range of individual variation, specific discrepancies in the relative abundance of functional classes indicate post-LECA adaptations (**Extended Data Fig. 1**). Notably, all trophic groups have significantly expanded repertoires of signal transduction proteins, suggesting post-LECA diversification in ecological and environmental interactions.

Distilling the ancestries of the LECA proteome

We next investigated the topologies of the reconstructed trees for each OG and inferred the phylogenetic sister group as the best proxy for ancestry (**see Methods, Fig. 2a, Supplementary Figure 3**). A significant fraction of OGs (41-43% depending on the dataset and stringency) had only eukaryotic homologs passing our filters, and we classified them as putative innovations. Considering only innovations predicted in all three datasets, and passing the strict criterion in at least one dataset we compiled a list of 1,188 KOs that could be considered Eukaryotic Signature Proteins (ESPs) **(Supplementary Table 8)**, which function in various pathways including transport and catabolism, signal transduction, translation, transcription, and folding, sorting and degradation processes. For OGs having non-eukaryotic homologs, we found a diversity of taxonomic assignments, in line with earlier findings[3]. The overall representation of major taxonomic groups remained highly similar across datasets and stringency criteria, indicating this result is independent from the eTOLDBs composition (**Supplementary Figure 3**).

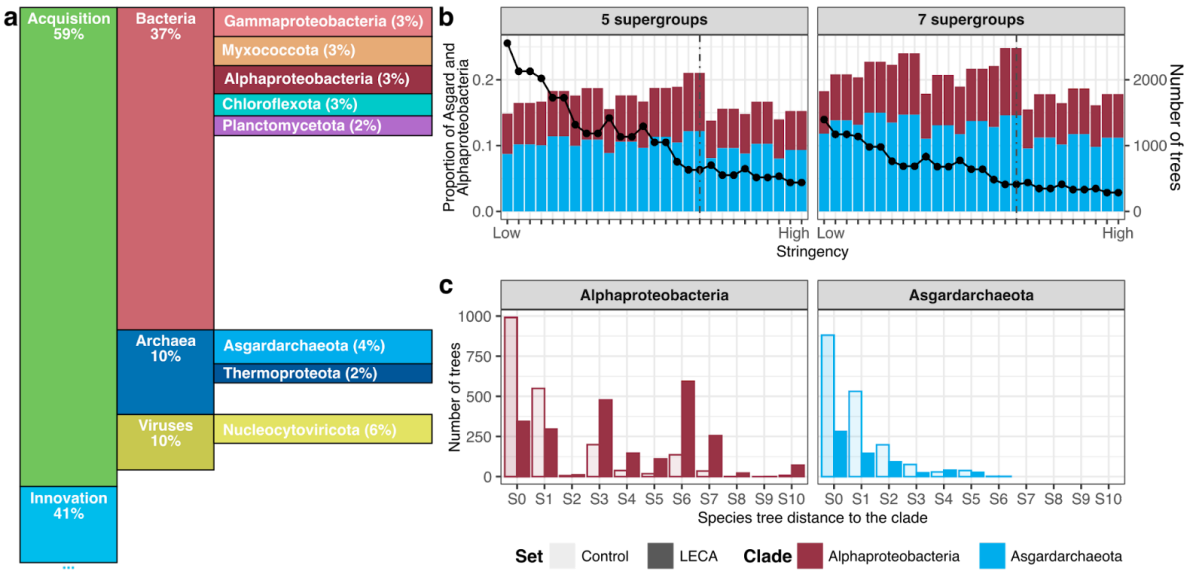

***Figure 2. Summary of the LECA ancestries**. a) Distribution of the ancestries of the LECA proteome. The consensus proportions have been calculated by the mean of the proportions in the three databases under the strict criterion, the innovation share has been shortened for visualisation purposes. b) Stress test for the number of trees and the relative contribution of Alphaproteobacteria and Asgardarchaeota to LECA through different stringency parameters: 5 and 7 eukaryotic supergroups for the definition of LECA, the stringency accounts for the UltraFast Bootstrap value in the branch connecting the non-eukaryotic sister with the LECA node, the taxonomic bootstrap for the definition of the donor and the proportion of the assigned donor in the first sister. The x-axis is sorted from less to more stringency. The*

vertical dashed line shows the threshold used to select the main non-eukaryotic contributors to LECA. c) Verticality test of the Alphaproteobacteria and Asgardarchaeota families. The lighter bar shows the number of trees that have the indicated species tree sister as sister to the control family. The darker bar shows the number of mLECA-OGs that have clades of that given sister as donors. The decreasing trend and correlation between the lighter and the darker bars would show that the genes evolve vertically.

To assess whether these alternative signals could result from phylogenetic noise, we performed a "stress test" across varying levels of stringency (**see Methods**). In doing so, we compared how the representation of alternative sister groups varied in relation to that of the accepted ancestors (Asgard archaea and Alphaproteobacteria) used here as "positive" controls (**Fig. 2b**). Increasing phylogenetic stringency reduced dramatically the number of trees considered, but only slightly varied the total proportion of OGs mapped to Asgard archaea and Alphaproteobacteria (**Fig. 2b**). In contrast, altering the number of required supergroups significantly affected the relative proportion of these two groups in favour of Asgard archaea, consistent with the core housekeeping functions of genes with this origin. Several taxonomic assignments showed similar support as the positive controls, and we selected those contributing at least 15 highly supported genes in any of the datasets (present in seven eukaryotic supergroups and with ultrafast and taxonomic bootstrap support values higher than 95%). These were the bacterial phyla (*sensu* GTDB) Gammaproteobacteria (which now includes betaproteobacteria), Myxococcota (including some members of former deltaproteobacteria), Alphaproteobacteria, Chloroflexota, and Planctomycetota; the archaeal clades Asgard archaea and Thermoproteota (comprising the TACK superphylum) and Nucleocytoviricota (a viral clade which includes giant viruses). We found no significant differences in terms of the amount of phylogenetic signal among sets of trees with different taxonomic assignments[21] (**Supplementary Figure 4)**. From these analyses, we conclude that phylogenetic noise alone cannot result in the observed diversity of taxonomic assignment patterns.

To assess what could be the expected level of variability of phylogenetic signal in genes inherited from the Alphaproteobacterial or Asgard archaeal ancestors, we applied our OG reconstruction and phylogenetic assignment procedure to Alphaproteobacterial and Asgard archaeal pangenomes, while removing these groups, respectively, from the BroaDB (**see Methods**). Expectedly, these "negative controls" show a majority of trees assigned to the closest sister groups of Alphaproteobacteria and Asgard archaea, respectively, followed by alternative taxonomic assignments with abundances that decrease as they correspond to more distant relatives (**Fig. 2c**). These results are compatible with a majority of the phylogenetic signal being driven by vertical inheritance coupled with gene loss. A similar pattern is observed for the LECA genes assigned to archaeal genes (**Fig. 2c**), suggesting that alternative non-asgard signals from archaea are likely the result of vertical inheritance through the Asgard ancestor coupled to loss in the sampled Asgards. This is in stark contrast to the observed relationships of LECA genes with bacterial ancestors **(Fig. 2c)**, indicating ancestries that cannot be explained by vertical inheritance through the mitochondrial ancestor. These analyses indicate that the diversity in taxonomic assignments is a composite of vertical and

horizontal inheritance, and suggest the existence of transfers from bacterial donors distinct from the mitochondrial ancestor to the protoeukaryote. From these results, we conclude that the most plausible explanation of the Thermoproteota signal results from loss or incomplete sampling of Asgard archaea. We conclude the same for the Gammaproteobacterial signal with respect to the Alphaproteobacteria. Recent studies place mitochondria as a sister group to, rather than within, Alphaproteobacteria[22], supporting a higher mixed signal in the Alpha/Gammaproteobacteria pair as compared to Asgard/Thermoproteota. For the remaining four groups, our results suggest they originated from viral or bacterial donors distinct from the ancestor of mitochondria.

Inferring plausible metabolisms in the putative prokaryotic donors

To investigate whether genes showing different ancestries might have been acquired in the form of "transfer waves" (i.e. when the inferred timings of different HGTs cluster around a similar period), we applied a previously described branch length ratio approach[3], as recently implemented into a Bayesian inference framework[23]. This approach measures lengths of the branch separating the ancestor of the LECA group with the most recent common ancestor with its non-eukaryotic sister group, normalised with the median branch length within LECA, which refers to the same geological time (from LECA to present) and therefore provides a relative time for the acquisition (**see Methods**). In this framework longer lengths are interpreted as more ancient relatedness. Our results (**Fig. 3b, Extended Data Fig. 2**) depict distinct branch length distributions peaking at different values, suggesting different relative timings. Although the specific relative ordering of the distributions vary across datasets and stringency levels, several consistent trends can be recognized. First, Thermoproteota and Asgard archaea are the most ancestral signals and Alphaproteobacteria tends to be the most recent. Note that the longer lengths in Thermoproteota genes with respect to Asgard archaea are compatible with a vertical origin of the former through the Asgard archaeal ancestor, as concluded above. As for Gammaproteobacteria, the branch lengths are very close to Alphaproteobacteria and overlap significantly, suggesting that the mitochondrial donor branched in between these two lineages. Among the alternative donors, Planctomycetes tend to be the one with the longer stem lengths, suggesting an early interaction, followed by an intermediate period with multiple contributions (Choroflexota, Myxococcota) in an uncertain order (**Extended Data Fig. 2**), and followed by the most recent relationship with the mitochondrial ancestor (**Fig. 3a**). The position of Nucleoviridycota is more uncertain, with an intermediate or late placement depending on the dataset. Notably, all identified donor bacterial taxa are common components of microbial mats, and the inferred order of acquisition of Asgard archaea, Planctomycetota and Proteobacteria is reminiscent of the relative depth in which each of these taxa is found in this type of environments[24] (**Fig. 3b**).

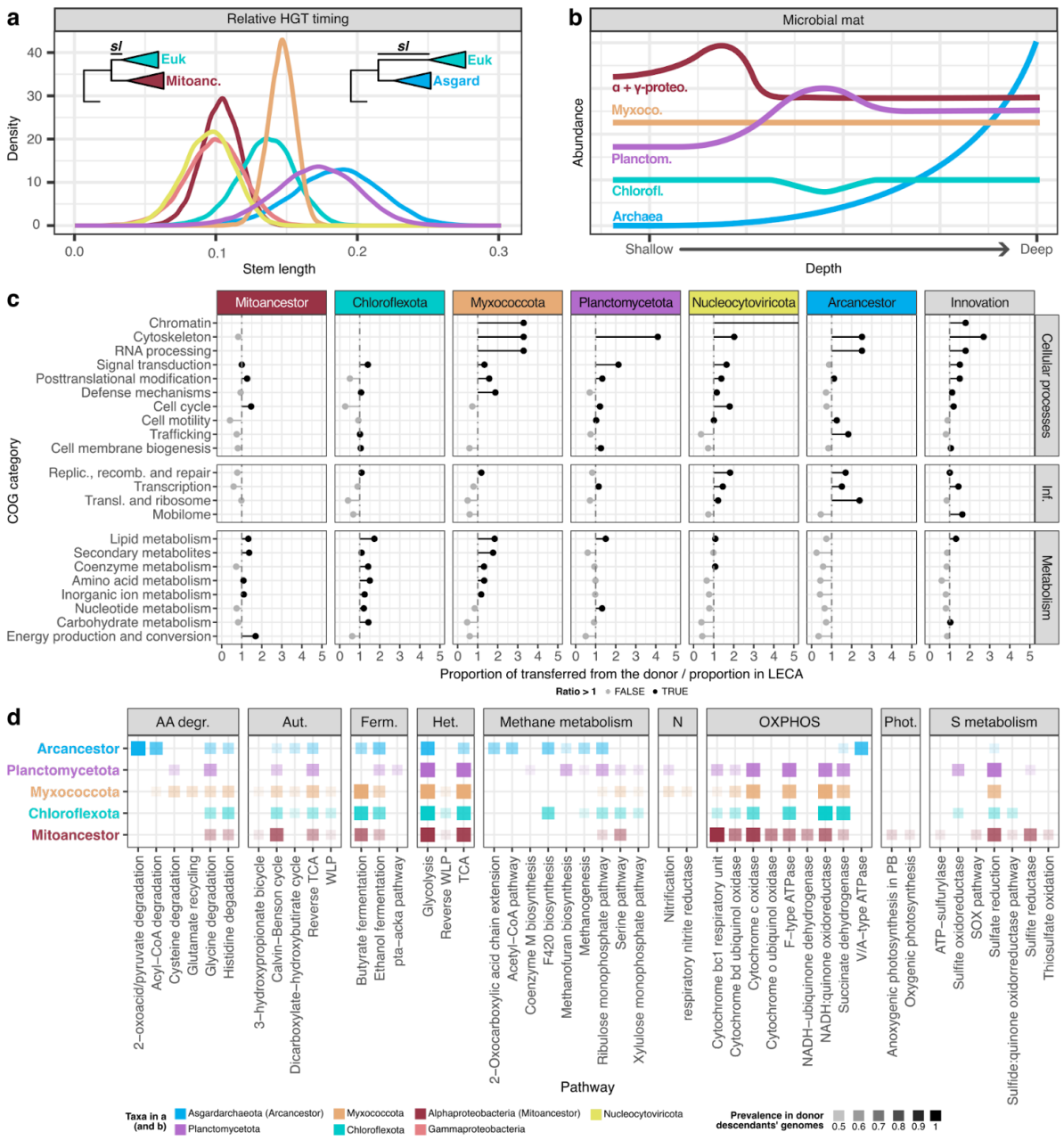

Figure 3. Characterisation of the putative prokaryotic contributors. *a) Posterior distribution of the mode for the normalised stem lengths of the non-eukaryotic acquisitions in LECA. Shorter stem lengths would refer to more recent transfers while longer would indicate earlier acquisitions. The colours indicate the different donors to LECA according to the legend at the bottom. The posterior distributions for each database and criterion are available in the* ***Extended Data Fig. 3****. b) Schematic relative abundances of the inferred donors across the different layers of a microbial mat according to*[24]*. c) Functional contributions of the non-eukaryotic genes to the LECA consensus proteome. Overrepresentation of COG categories among the transferred genes to LECA in contrast to the whole LECA proteome. Values above 1 mean that the gene is more frequent among the donor's contributions than in LECA. d) Prevalence of some metabolic features for each donor. The colour indicates the donor according to the legend. The opacity of the colour is the proportion of genomes with at least 50% of the metabolic pathway completeness.*

Pathways and KOs with more than 0.4 prevalence are shown. A detailed version of this plot, with the prevalence for all the metabolic pathways and KOs assessed, is available in the ***Extended Data Fig. 4***.

We used these branch length distributions to refine the set of proteins likely acquired from the same donor in a single wave by removing protein families with anomalous branch lengths (**see Methods**). Given the considerations explained above, we merged genes from Thermoproteota and Asgard assignments, as well as Gamma- and Alphaproteobacteria. We first assessed whether these transfers were enriched in specific functions (**Fig. 3c**). Asgard contributions (including those mapped to Thermoproteota) are involved in informational functions (transcription and translation processes and nucleic acid pathways, (**Fig. 3c**). Conversely, bacterial contributions are enriched in metabolic functions. Alphaproteobacteria (including genes mapped to Gammaproteobacteria) transferred core functions of the carbon and energy metabolism as well as the metabolism of lipids and secondary metabolites. Alternative bacterial donors also have a distinct signature in the LECA proteome: proteins of Chloroflexota origin are enriched in all the metabolic COG categories except energy production and conversion, whereas those coming from Myxococcota are enriched in cellular processes such as chromatin, cytoskeleton and RNA processing. Planctomycetota contributions have more genes involved in the metabolism of lipids. The viral contribution to LECA is enriched in signalling and cell cycle control functions, which are complemented with innovations. Many other functions are enriched among innovations, such as the cytoskeleton, as well as categories involved in cellular processes, which complement the acquisitions from Asgard archaea.

Then, we inferred the metabolic capacities possibly present in each considered donor beyond the genes transferred to LECA. For this, we retrieved genomes of the donor clade that intersected with at least 50% of the KOs transferred to LECA by this donor and reconstructed the potential metabolic properties of each donor, using the information of the KO and metabolic modules prevalence in this subset (**see Methods**). These reconstructions, which expand KO acquisitions in LECA with commonly co-occurring KOs in extant organisms of that donor clade, provide a first approximation for the core set of metabolic pathways possibly present in the ancestral donors (**Extended Data Fig. 4, Supplementary Data**).

Taken together, these observations depict a complex scenario involving a series of HGT waves from a diverse set of donors into the proto-eukaryotic lineage, as proposed by the pre-mitochondrial symbioses hypothesis[5]. Notably, the inferred taxonomic identity, metabolic features, and order of acquisition also support previous ideas of syntrophic origins within microbial mats[25] (**Fig. 3b**), but also provide remarkable new insights. Although the HS syntrophy hypothesis proposed Myxococcota as the host of the eukaryogenesis[4], the presence of Planctomycetota as one of the earliest contributors in our analysis, and their notable cellular complexity[26], including the potential ability for cell engulfment[27], opens the door to consider Plantomycetota a potential alternative host, a possibility that had been earlier proposed[28,29], but later criticised due the lack of evidence for this origin in LECA proteomes[30]. As to the potential metabolic interactions, the incomplete nature of our

reconstructions makes it difficult to favour a specific proposal, but provide information to contrast different proposed models. For instance, our results strongly support both a methanogenic nature of the Asgard archaea as proposed in various models[4,25,31], and a fermentative metabolism as suggested in the reverse-flow model[32]. Sulphur reduction is inferred for the Myxococcota donor as proposed in the revised HS syntrophy hypothesis[4] but also in the Planctomycetota donor, suggesting the model could be adjusted to this alternative host to accommodate our relative timing observations. Both cultured Asgard archaea grow in syntrophy with sulfate-reducing Desulfobacteriota partners[33,34], and sulfur-based interactions have been hypothesized between extant Asgard archaea and Planctomycetota in mangrove sediments[35], suggesting this type of syntrophy can be established between Asgard archaea and diverse bacterial groups. We also observe a weak potential for ammonium oxidising (anammox) metabolism in the Planctomycetota donor, and for Nitrogen fixation in the mitochondrial ancestor. Importantly, the reconstruction of the potential metabolic features present in the mitochondrial ancestor strongly supports the presence of aerobic respiration but lend no support to the presence of Fe-Fe- Ni-Fe- and Fe-only hydrogenases as implied by the Hydrogen hypothesis or the reverse-flow models[25,31,32]. Nevertheless, further phylogenetic, genomic, ecological and biochemical research is needed to elucidate the nature of these early interactions[30].

Contributions from Nucleocytoviricota

The contribution of viruses in shaping the proteome of LECA has been discussed in different earlier studies[36–38]. Our analysis provides a systematic assessment of the potential early gene exchanges between viruses and pre-LECA eukaryotes, and underscores a large contribution of genes potentially exchanged with a specific type of viruses: Nucleocytoviricota, which are associated to ~9.4% of the acquired protein families in LECA. Nucleocytoviricota, also referred to as nucleocytoplasmic large DNA viruses, is a phylum of viruses known to infect diverse eukaryotes and to comprise giant viruses with large genomes[39]. We explored trees that showed Nucleocytoviricota as the direct sister to LECA and that were in the same acquisition wave. 226 out of the 505 genes (45%) had a bacterial sister that preceded the divergence between LECA and the viral taxonomic groups, suggesting the DNA viruses may have mediated gene transfer into the LECA lineage (**see Supplementary Discussion**). These bacterial sisters contained, among other bacterial groups, all of the donor groups detected in the previous analysis. Functional analysis of viral-transferred genes indicated over-abundance of genes related to signal transduction mechanisms, mainly kinases, and post-translational modification proteins. This observation is compatible with other findings indicating this type of viruses infected pre-LECA ancestors[36], and suggest they may have facilitated the transfer from interacting organisms, including the major donors identified here and possibly other coexisting eukaryote-like lineages.

Discussion

Our results confirm and expand earlier results supporting significant gene flow from diverse prokaryotic ancestors preceding LECA[3,40]. Considering the distributions of phylogenetic

distances to the closest relatives, our data support pre-mitochondrial waves of gene acquisitions from additional major prokaryotic donors as the most likely scenario[5,41], against a backdrop of continuous HGT contributions from a diversity of other minor sources. Although the number of significant donors varies depending on the chosen threshold, we find support for at least three non-Alphaproteobacterial bacterial donors transferring a significant amount of genes: Planctomycetota, Chloroflexota, and Myxococcota. Transfers from these non-Alphaproteobacterial donors have been identified in earlier studies[3,18,40], including small-scale detailed ones such as the origin of some steroid biosynthesis enzymes from Myxococcota[42,43]. Our results also underscore a central contribution from Nucleocytoviricota, and suggest a potential role of these viruses in mediating some of the gene transfers from the bacterial donors.

Taken together, our data do not fit current models of eukaryogenesis involving a single interaction between an Asgard archaeal ancestor and an Alphaproteobacterial mitochondrial ancestor, and suggest they should be expanded to cover additional interactions with other relevant partners. Our relative timing estimates suggest early interactions between an Asgard archaea and a Plantomycetota, followed by contributions from Chloroflexota and Myxococcota, and an intermediate or late acquisition of the mitochondrion. Thus, our results indicate that the prokaryotic-to-eukaryotic transition was a gradual and complex process, with a constant flow of HGT from diverse minor and major sources that would have required successive, but perhaps overlapping, tight ecological interactions between different bacterial partners, each contributing a significant set of genes and functions to the pre-LECA ancestor.

The significant amount and diversity of transferred genes from each of the considered major donors is unlikely without some form of sustained ecological interaction, but our data do not inform on the type of interaction (endosymbiotic, ectosymbiotic, parasitic, etc). Nevertheless, the inferred metabolic capabilities of the inferred donors can be used to contrast with existing metabolic scenarios, to adjust them, and to propose novel ones. Microbes are known to form complex communities and closely interact as part of structures such as microbial mats or complex biofilms[44], and it is reasonable to consider that the ancestors of LECA lived in such rich and complex environments. The presence of transporters, junction proteins, and quorum sensing systems in our LECA reconstruction indicates a high interaction potential. Of note, the relative ordering of gene acquisitions parallels that of microbial mats, with archaea being more abundant in deeper anoxygenic layers, in intermediate layers the abundance of Planctomycetes is higher and upper layers are enriched in Proteobacterial sequences[24]. Our data also suggests that viruses were also an active part of this ecosystem and that they possibly mediated some of the gene transfers. This also raises the possibility that different coexisting proto-eukaryotic lineages could also have exchanged genes in this manner, thereby facilitating adaptation.

Altogether, our results challenge models depicting eukaryogenesis as a one-off event resulting from a unique symbiotic interaction between just two prokaryotes, and support a paradigm shift towards more gradual scenarios in which the prokaryotic-to-eukaryotic transition was achieved through serial symbiotic interactions with a diverse range of bacterial

partners leading to significant gene transfer, some of which were possibly mediated by viruses.

Main references

1. López-García, P. & Moreira, D. Open Questions on the Origin of Eukaryotes. *Trends Ecol. Evol.* **30**, 697–708 (2015).
2. Gabaldón, T. Origin and Early Evolution of the Eukaryotic Cell. *Annu. Rev. Microbiol.* **75**, 631–647 (2021).
3. Pittis, A. A. & Gabaldón, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).
4. López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).
5. Gabaldón, T. Relative timing of mitochondrial endosymbiosis and the 'pre-mitochondrial symbioses' hypothesis. *IUBMB Life* **70**, 1188–1196 (2018).
6. Forterre, P. & Prangishvili, D. The major role of viruses in cellular evolution: facts and hypotheses. *Curr. Opin. Virol.* **3**, 558–565 (2013).
7. Irwin, N. A. T., Pittis, A. A., Richards, T. A. & Keeling, P. J. Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. *Nat Microbiol* **7**, 327–336 (2022).
8. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 19585–19592 (2019).
9. Stanier, R. Y. *The Microbial World: By Roger Y. Stanier, Michael Doudoroff and Edward A. Adelberg*. (1963).
10. Gray, M. W. Lynn Margulis and the endosymbiont hypothesis: 50 years later. *Mol. Biol. Cell* **28**, 1285–1287 (2017).
11. Imachi, H. *et al.* Promethearchaeum syntrophicum gen. nov., sp. nov., an anaerobic, obligately syntrophic archaeon, the first isolate of the lineage 'Asgard' archaea, and proposal of the new archaeal phylum Promethearchaeota phyl. nov. and kingdom Promethearchaeati regn. nov. *Int. J. Syst. Evol. Microbiol.* **74**, (2024).
12. Eme, L. *et al.* Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).
13. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
14. Gabaldón, T. & Huynen, M. A. Reconstruction of the proto-mitochondrial metabolism. *Science* **301**, 609 (2003).
15. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–105 (2018).
16. Gabaldón, T. & Huynen, M. A. From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. *PLoS Comput. Biol.* **3**, e219 (2007).
17. Pisani, D., Cotton, J. A. & McInerney, J. O. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* **24**, 1752–1760 (2007).

18. 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).
19. 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).
20. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A. S. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* **3**, (2018).
21. Haag, J., Höhler, D., Bettisworth, B. & Stamatakis, A. From Easy to Hopeless-Predicting the Difficulty of Phylogenetic Analyses. *Mol. Biol. Evol.* **39**, (2022).
22. Muñoz-Gómez, S. A. *et al.* Site-and-branch-heterogeneous analyses of an expanded dataset favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* **6**, 253–262 (2022).
23. Bernabeu, M., Armero, C. & Gabaldón, T. Probabilistic modelling improves relative dating from gene phylogenies. *bioRxiv* 2024.09.30.615760 (2024) doi:10.1101/2024.09.30.615760.
24. Gutiérrez-Preciado, A. *et al.* Functional shifts in microbial mats recapitulate early Earth metabolic transitions. *Nat Ecol Evol* **2**, 1700–1708 (2018).
25. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**, 37–41 (1998).
26. Boedeker, C. *et al.* Determining the bacterial cell biology of Planctomycetes. *Nat. Commun.* **8**, 14853 (2017).
27. Wurzbacher, C. E. *et al.* 'Candidatus Uabimicrobium helgolandensis'-a planctomycetal bacterium with phagocytosis-like prey cell engulfment, surface-dependent motility, and cell division. *MBio* e0204424 (2024).
28. Devos, D. P. & Reynaud, E. G. Evolution. Intermediate steps. *Science* **330**, 1187–1188 (2010).
29. Devos, D. P. Reconciling Asgardarchaeota Phylogenetic Proximity to Eukaryotes and Planctomycetes Cellular Features in the Evolution of Life. *Mol. Biol. Evol.* **38**, 3531–3542 (2021).
30. McInerney, J. O. *et al.* Planctomycetes and eukaryotes: a case of analogy not homology. *Bioessays* **33**, 810–817 (2011).
31. Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N. & Martin, W. F. Lokiarchaeon is hydrogen dependent. *Nat Microbiol* **1**, 16034 (2016).
32. 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).
33. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
34. Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332–339 (2023).
35. Du, H. *et al.* Microbial active functional modules derived from network analysis and metabolic interactions decipher the complex microbiome assembly in mangrove

sediments. *Microbiome* **10**, 224 (2022).
36. Krupovic, M., Dolja, V. V. & Koonin, E. V. The virome of the last eukaryotic common ancestor and eukaryogenesis. *Nat Microbiol* **8**, 1008–1017 (2023).
37. Bell, P. J. L. Evidence supporting a viral origin of the eukaryotic nucleus. *Virus Res.* **289**, 198168 (2020).
38. Forterre, P. & Gaïa, M. Giant viruses and the origin of modern eukaryotes. *Curr. Opin. Microbiol.* **31**, 44–49 (2016).
39. Aylward, F. O., Moniruzzaman, M., Ha, A. D. & Koonin, E. V. A phylogenomic framework for charting the diversity and evolution of giant viruses. *PLoS Biol.* **19**, e3001430 (2021).
40. Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol* **5**, 92–100 (2021).
41. Bernabeu, M., Manzano-Morales, S. & Gabaldón, T. On the impact of incomplete taxon sampling on the relative timing of gene transfer events. *PLoS Biol.* **22**, e3002460 (2024).
42. 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).
43. Hoshino, Y. & Gaucher, E. A. Evolution of bacterial steroid biosynthesis and its impact on eukaryogenesis. *Proceedings of the National Academy of Sciences* **118**, e2101276118 (2021).
44. Bolhuis, H., Cretoiu, M. S. & Stal, L. J. Molecular ecology of microbial mats. *FEMS Microbiol. Ecol.* **90**, 335–350 (2014).

Figure legends

Figure 1. Reconstruction of the ancestral proteome of LECA and its metabolic features. a) Metabolic reconstruction of the consensus LECA proteome. In light grey: the metabolic map of KEGG. In colours: the modules inferred to have been present in LECA using the consensus proteome, which accounts for the KOs present in at least 2 LECA reconstructions of different eTOLDBs or supported by a LECA group with five eukaryotic supergroups. Each colour refers to a metabolic category according to the coloured boxes. b) Frequency of COG database functional categories in the consensus LECA proteome (black) and in the 63 FLUPs (distribution in blue). c) Percentage of KEGG orthologs (KOs) shared between the ancestral proteome of LECA and extant free-living unicellular autotrophs, osmotrophs and phagotrophs (FLUAs, FLUOs and FLUPs, respectively). d) Relationship between the percentage of the proteome inherited from LECA and the extant genome size.

Figure 2. Summary of the LECA ancestries. a) Distribution of the ancestries of the LECA proteome. The consensus proportions have been calculated by the mean of the proportions in the three databases under the strict criterion, the innovation share has been shortened for visualisation purposes. b) Stress test for the number of trees and the relative contribution of Alphaproteobacteria and Asgard archaea to LECA through different stringency parameters: 5 and 7 eukaryotic supergroups for the definition of LECA, the stringency accounts for the UltraFast Bootstrap value in the branch connecting the non-eukaryotic sister with the LECA node, the taxonomic bootstrap for the definition of the donor and the proportion of the assigned donor in the first sister. The x-axis is sorted from less to more stringency. The vertical dashed line shows the threshold used to select the main non-eukaryotic contributors to LECA. c) Verticality test of the Alphaproteobacteria and Asgard archaea families. The lighter bar shows the number of trees that have the indicated species tree sister as sister to the control family. The darker bar shows the number of mLECA-OGs that have clades of that given sister as donors. The decreasing trend and correlation between the lighter and the darker bars would show that the genes evolve vertically.

Figure 3. Characterisation of the putative prokaryotic contributors. a) Posterior distribution of the mode for the normalised stem lengths of the non-eukaryotic acquisitions in LECA. Shorter stem lengths would refer to more recent transfers while longer would indicate earlier acquisitions. The colours indicate the different donors to LECA according to the legend at the bottom. The posterior distributions for each database and criterion are available in the **Extended Data Fig. 3**. b) Schematic relative abundances of the inferred donors across the different layers of a microbial mat according to[24]. c) Functional contributions of the non-eukaryotic genes to the LECA consensus proteome. Overrepresentation of COG categories among the transferred genes to LECA in contrast to the whole LECA proteome. Values above 1 mean that the gene is more frequent among the donor's contributions than in LECA. d) Prevalence of some metabolic features for each donor. The colour indicates the donor according to the legend. The opacity of the colour is the proportion of genomes with at least 50% of the metabolic pathway completeness. Pathways and KOs with more than 0.4 prevalence are shown. A detailed version of this plot, with the prevalence for all the metabolic pathways and KOs assessed, is available in the **Extended Data Fig. 4**.

Methods

1. Data retrieval, curation and proteome database reconstruction

1.1. Eukaryotic Tree of Life database (eTOLDB) construction

To guide dataset selection and LECA reconstruction, we used the phylogenetic backbone of the eukaryotic tree of life (eTOL) and the nine eukaryotic supergroups classification depicted in (**Supplementary Figure 1**). This is a consensus between recent eTOL reconstructions[45–48] and an operational definition of two stems and nine supergroups that largely reflect previous classifications[49]. Additional details are provided in (**Supplementary Methods**).

We manually selected and retrieved 276 eukaryotic proteomes (**Supplementary Table 1**) downloaded from NCBI[50], EukProt[51], P10K[52] and UniProt[53], Ensembl[54] and SGD[55], aiming to cover all supergroups represented in our reference phylogenetic backbone (**Supplementary Figure 1**). We aimed at a balanced representation of the divisions and prioritised the most complete and highest quality genomes for each considered division based on information available in the hosting databases. For each proteome, we only kept the longest isoform per gene, as indicated in the GFF files or FASTA headers. The isoform-free proteomes were then submitted to an analysis of completeness using BUSCO v5.5[56] with default parameters and the eukaryotic_odb10, and we extracted basic statistics using seqkit stats v2.8.2[57]. To assess for the presence of uncertain amino acids (Xs) and low complexity regions in the proteomes, we ran SEG[58] with default parameters. We used an in-house script (See “Data availability”) to trim the low-complexity ends of the proteins and calculated the coverage of low-complexity regions in the protein. We considered a low-complexity protein if the coverage of low-complexity regions was higher than 25% of the protein and discarded it for the downstream analyses. Moreover, we calculated the proportion of low-complexity proteins in the whole proteome as a measure of the quality of the protein set. We also discarded proteins that exceeded the 10,000 amino acids or were smaller than 50 amino acids. This resulted in a set of clean proteomes. To remove redundancy and recent (with respect to eukaryotic evolution) duplications, we clustered each proteome individually using MMseqs v2[59] with strict thresholds: minimum amino acid identity percentage of 80%, minimum target coverage of 0.5 and coverage mode 2 (--min-seq-id 0.8 -c 0.5 --cov-mode 2). The representative sequences of the clusters constituted the final proteome. To assess the effect of the cleaning and clustering steps, we ran again BUSCO v5.5 and seqkit stats v2.8.2. We discarded the proteomes that suffered a combined reduction in the number of proteins higher than 50%, having a low-complexity proteins percentage higher than 50% and a BUSCO completeness reduction higher than 30%. This resulted in 30 proteomes being excluded, and we continued with the 256 remaining proteomes.

We built three eTOL datasets of 100 proteomes each, which we named eTOLDBA, eTOLDB and eTOLDBC. For this, we first classified the 256 proteomes based on taxonomy and distributed them in nine different supergroups and 28 divisions (**Supplementary Figure 1**). When less than four species were available for a division, we included all in the three eTOLDB databases. For the remaining divisions we selected different, partially overlapping, species sets for each eTOLDB. eTOLDBA contained, whenever possible, the most complete

proteomes and the fewest parasites. Proteomes for model species *Homo sapiens, Arabidopsis thaliana* and *Saccharomyces cerevisiae* were included in the three databases despite being part of large taxonomic groups. This was done for later annotation purposes. The list of chosen species for each dataset can be found in **Supplementary Table 1**. Finally, a total of 185 species were present in the three eTOLDBs with an overlap of 58 species between eTOLDBA and eTOLDBB, and 53 species between eTOLDBA - eTOLDBC and eTOLDBB - eTOLDBC. This resulted in an average protein overlap of 46%.

1.2. Broad prokaryotic database construction (BroadDB)

To ensure an even sampling of the prokaryotic genome repertoire, while minimising the impact of intra-domain HGT as a confounding factor of prokaryotic ancestry, we downsampled available prokaryotic genomes using a pangenome approach. For this, we took all species representatives in GTDB r207[19] (62,291 bacteria and 3,412 archaea) and selected one genome per genus (*sensu* GTDB), choosing the one with the maximal CheckM[60] completeness and, in case of ties, minimal CheckM contamination. The proteomes of each genus representative were then grouped at order level and clustered with mmseqs2 easy-linclust[59] (--min-seq-id 0.3 -c 0.5 --cov-mode 0 --cluster-mode 2 -e 0.001). We kept a sequence representative for clusters present in 15% or more of genus representatives in that order, as 15% is the cut-off criteria for the "cloud" component of the pangenome[61] (Supplementary figure cutoffs). To this dataset we added representative sequences from 1,319,927 viral protein clusters available at RVDB v25.0[20]. For taxonomic assignment of the sequences, we follow the GTDB and RVDB taxonomies throughout the manuscript.

2. Reconstruction of LECA

2.1. LECA-Orthogroup (LECA-OG) reconstruction

We ran OrthoFinder v2.5.5[62], using BLAST v2.15.0+[63] for the homology search with the filter for low complexity (-seq yes). Before the clustering step, BLAST results were filtered to remove hits with an e-value above 1e-05 and an overlap below 0.5. We additionally removed hits with disparate sizes (when hit was >2.5 longer or <0.5 shorter than the query) to avoid artefactual clustering of families due to domain sharing. The clustering part of OrthoFinder was executed using two different inflation parameters (I = 1.5 and I = 3.0). Little difference was found between the two predictions regarding number of LECA groups and consistency therefore an inflation parameter of 1.5 was chosen to go forward.

Orthogroups as defined by OrthoFinder were filtered to select those likely present in LECA (LECA-OG) based on their distribution across the eTOL. LECA-OGs had to comprise proteins from at least five species, and have representatives of the two stems and three or more different supergroups for a relaxed definition of LECA and five or more supergroups for a strict definition of LECA. These same criteria were used in downstream analyses.

2.2. Expanding LECA-OGs with closest non-eukaryotic homologs

To identify closest non-eukaryotic homologs and define the putative origins of LECA-OGs, we built a sequence profile for each of them. We aligned the sequences using MAFFT v7.525[64] and evaluated the completeness score (i.e. the percentage of unambiguous and non-gappy residues of the aligned sequence). We removed sequences with a completeness score lower than 75%, unless this removed more than 10% of the sequences in the alignment, in which case we removed the 10% of the sequences among those with the lowest completeness scores.

We re-aligned the remaining sequences with MAFFT v7.525[64] and built a profile with HMMbuild v3.3.2[65]. We ran hmmsearch v3.3.2 with this profile against the BroadDB described above. Results were filtered using an e-value threshold of 0.1 for both the complete sequence and the first domain found. Then, to ensure that the coverage requirements were maintained we performed a BLASTP v2.15.0+[63] search for each resulting prokaryotic sequence to the database of Eukaryotic sequences from the orthogroup. BLAST results were filtered with an e-value threshold of 1e-05, a coverage threshold of 0.5 over the full length of the prokaryotic sequence and ensuring that the target sequence was not 2.5 times longer or 2.5 times shorter than the query sequence. The top-hits that passed these filters were considered the closest prokaryotic homologs and were added to the sequences of the corresponding LECA-OG to compute phylogenies (see section below). The number of prokaryotic sequences in the trees was limited in the following way: if there were less than 400 eukaryotic sequences then prokaryotic sequences were added until reaching the 500 sequences. When more than 400 eukaryotic sequences were present, up to 100 prokaryotic sequences were added. Eukaryotic groups with more than 500 sequences processed differently. First genes were assigned a KO (see below) and then sequences from the same OG with the same KO were grouped together and processed as separate OGs.

2.3. Phylogenetic tree reconstructions

We used the PhylomeDB tree reconstruction pipeline[66] with slight modifications. In brief, we computed a consensus alignment using MergeAlign v3[67] from the alignments obtained by three different alignment programs (MUSCLE v5.2[68], MAFFT v7.525[64] and Kalign v3.4.0[69]) in forward and reverse orientations. This alignment was then trimmed using trimAl v1.4.rev15[70] with the -gappyout option. Finally, a maximum likelihood (ML) tree was reconstructed using IQ-TREE v2.3.4[71] and parameters -wbtl -B 1000 -m TEST -mset WAG,LG,JTT --madd LG4X,LG4M --mrate ALL. We used ModelFinder[72] to select the fittest substitution model among a subset of the available models in IQ-TREE. We selected the single matrix substitution models DCmut[73], JTTDCMut[73], VT[74], WAG[75] and LG[76] and the mixture models, LG4X and LG4M[77]. Branch support was assessed with 1,000 Ultra Fast Bootstrap replicates[78].

2.4. Monophyletic LECA orthogroups (mLECA-OG)

For each LECA-OG, we reconstructed a phylogeny following the phylogenetic reconstruction pipeline detailed above. We annotated the tree with the taxonomy of each group and retrieved the largest monophyletic clade containing only eukaryotic sequences and fulfilling the

relaxed LECA criterion (5 species, 3 supergroups, and 2 stems, see above). This was considered the "starting LECA node". We rooted the phylogeny in the non-eukaryotic node farthest to the starting LECA node. We next iterated through the sister clades of the starting LECA node and incorporated them into the LECA clade if:

1. The sister was mostly eukaryotic: that is the fraction of non-eukaryotic sequences was lower than or equal to 0.5, for sister clades with up to 20 sequences, or lower than or equal to 0.25 if the sister had more than 20 tips.
2. In the case that the first sister was mainly non-eukaryotic, but the second sister was mostly eukaryotic (according to the criteria above), the first sister was considered a transfer from the eukaryotic clade. In this case, we incorporated the first sister if all the following conditions were met:
    a. the first sister had a number of sequences lower than 75% of the number of sequences in the LECA clade
    b. the number of non-eukaryotic phyla was lower than 3 (assuming low diversity of the first non-eukaryotic sister), and
    c. the fraction of non-eukaryotic sequences in the second sister is lower than 25%.

Once the LECA clade was defined with the above criteria, we removed it from the tree and performed the search iteratively until no more starting LECA nodes could be found in the tree.

For each LECA node, we selected only the eukaryotic sequences (mLECA-OG hereafter). Finally, we aligned these eukaryotic filtered sequences to compute a profile-based search of non-eukaryotic homologs using the above-mentioned criteria. We reconstructed the tree of the resulting set of eukaryotic and prokaryotic sequences using the phylogenetic reconstruction described above and identified LECA nodes in this tree following the same described procedure. This resulted in a final set of mLECA-OGs which is a refinement of the previous one. Only this final set was used in subsequent analyses.

2.5. Functional annotations of mLECA-OGs

We annotated all the proteins with KOfamScan v1.3.0 (https://github.com/takaram/kofam_scan)[79] and HMM search against the COG database. Results from KofamScan were filtered in the following way: First, we selected those hits considered significant by KofamScan. When multiple hits were found, we selected the hit with the lowest e-value and highest hmmsearch score. In addition, for proteins with no hits passing the significance threshold of KOfamScan, we selected the hit with lowest e-value and highest hmmsearch score. To benchmark the accuracy of our predictions, we used 36 proteomes in our dataset that had been obtained from Uniprot and had mappings to genes in KEGG. This allowed us to establish KOs for a total of 154,248 proteins, independent of KofamScan. For those proteins we calculated the average correspondence of KO predictions per each species. We obtained 95% correct predictions on average, validating our KO annotation method. We further assigned each protein to a COG family by performing an

hmmsearch using the COG profiles and selecting the first hit based on bitscore. We kept those COGs with a bit score higher than the percentile 90. Using the definitions table, we assigned the COG category to each protein.

Once all proteins were functionally annotated, we assigned a KO for each set of sequences (LECA-OGs, mLECA-OGs, or sister group) using the same approach of eggNOG orthogroups functional annotation, which takes into account the proportion of the KO within the set of sequences and its overrepresentation with respect to the background KOs distribution (the whole set of annotated proteins). We chose a single KO or multiple KOs in case of ties by minimising the sum of the KO rank in the proportions list and the overrepresentation list. To annotate the mLECA-OGs, we used the common KOs assigned to the LECA clade and its sister if they agreed, and the one of the LECA clade if they disagreed.

2.6. Inference of the LECA proteomes

For each eTOLDB version (A, B and C) and each LECA criterion (three supergroups and five supergroups), we obtained the set of mLECA-OG that fulfilled these criteria and stored them in tables accounting for their origins and KO annotation. These proteomes are the ones that we used to infer the general proportion of acquisitions and innovations, as well as the general metabolism of LECA. Furthermore, we calculated a consensus proteome with the KO annotations that were pervasive and well-supported across the datasets to make inferences about the LECA features and metabolism. For this consensus we considered KOs present in two out of three TOLDB datasets or if at least five supergroups supported that KO in any TOLDB dataset.

2.7. LECA metabolism and features inference

We inferred the presence of metabolic pathways and their completeness for each LECA proteome (one per eTOLDB version and LECA criterion) using the module anvi-estimate-metabolism of the Anvi'o v8[80] package. To represent the modules acquired from the different domains of life, we first retrieved the proportion of blocks of the pathway that originated from the same non-eukaryotic domain (Archaea, Bacteria or Viruses). Secondly, if the proportion of blocks in a module that originated from a specific domain was higher than 0.6, we assumed that the given pathway had a non-eukaryotic origin. We, finally, plotted the reconstructed metabolic map using ggKEGG v1.2.3[81].

For other molecular components that are not necessarily associated to a single KEGG map such as the complexes that form the flagellum, the spliceosome or the cytoskeleton, we collected the set of genes associated to these components as depicted by KEGG brite, complemented them based on literature when necessary, and then checked the assigned origins of the different genes.

2.8. Comparison with FLUPs, FLUOs and FLUAs

In order to benchmark the inferred LECA proteome, we compared it to the core and individual proteomes of selected unicellular, heterotrophic eukaryotic organisms. For this, we selected, from the eTOLDB proteomes and the P10K project[52], organisms that were

unicellular, heterotrophic free-living, and phagotrophs, autotrophs and osmotrophs (FLUPs, FLUAs and FLUOs, respectively, **Supplementary Table 6**). After cleaning and performing quality checks on the proteomes (see Methods 1.1), we functionally annotated the proteomes following the same protein annotation pipeline as for the eTOLDB genomes (see Methods 2.5). In the case of the phagotrophs, which are the most taxonomically diverse, we defined orthogroups with OrthoFinder v2.5.5[62] using the same parameters as for the LECA-OGs. We annotated the orthogroups following the mLECA-OGs annotation pipeline (see Methods 2.5), and compared them to the LECA consensus proteome.

Based on the KO predictions for each genome included in FLUPs, FLUOs and FLUAs we calculated the percentage of each genome that was shared with LECA in terms of number of KOs (number of shared KOs / number of KOs predicted in the eukaryotic proteome). We also obtained the frequency in which each COG functional category appeared in each genome within the three datasets as the number of genes that had a COG functional category divided by the number of genes with at least one COG category assigned.

3. Tracing mLECA-OG ancestries

3.1. Identifying the origins of the mLECA-OGs

To detect the non-eukaryotic groups that transferred a given gene family to the protoeukaryote, we analysed the first sister of the mLECA-OG. First, we discarded a non-eukaryotic origin for the mLECA-OGs for those trees where the sister clade contained more than 25% eukaryotic sequences, as although the sister would be mainly non-eukaryotic, the presence of eukaryotic sequences could be the result of secondary transfers from eukaryotes to a broad set of bacteria, hampering the identification of a confident sister. For the trees that passed this initial filter, we retrieved the proportion of each non-eukaryotic group present in the sister clade. Finally, to assign a taxon to the sister clade, we selected the most abundant group in the first sister (i.e., the one with the maximum proportion). In the case of more than one non-eukaryotic group with the maximum proportion in the first sister, we assumed that the clade was mixed and calculated its common ancestor, which could be Bacteria, Archaea, Viruses or a mixture of them.

Trees in which the mLECA-OG was assigned to Nucleocytoviricota were assessed in more detail to assess the directionality of the transfer or the presence of secondary transfers from other groups. For this, we scanned the trees to search for additional, non-viral sequences. Given each tree with a sister in Nucleocytoviricota we located the point where LECA and its closest sister were located and went back in the tree towards the root in order to locate the first sister where the largest signal did not belong to Viruses. When such cases were found, the taxonomy of that sister was assessed and considered as the potential source of genes transferred to LECA through the viral lineage. Trees with only viruses as sisters and no other prokaryotic signal in the tree were discarded as it was difficult to assess whether they were the result of a more recent HGT event.

3.2. "Stress test" for identifying the main non-eukaryotic contributors

To calibrate the parameters that we used for the identification of the main bacterial, archaeal and viral contributors to the genome of LECA, we assessed a set of variables that allows us to figure out the status of the certainty about both LECA and the donor. We assessed the congruence of LECA and its relationship to its sister by checking the ultrafast bootstrap support (UfBS) values of these branches. However, for the congruence of the taxonomic assignment of the sister, we calculated the taxonomic bootstrap. The evolutionary depth of this project required a bootstrap support measure that accounted for broader taxonomic patterns rather than specific sequences moving through different nodes. Thus, we designed the taxonomic bootstrap, a measure of the support of a given taxonomic affiliation for a certain node of the tree. This version of the bootstrap calculates the proportion of ultrafast bootstrap trees that support the same taxonomic assignment for the sister of the ML tree regardless of the congruence of the sequences. In this case, we analysed the taxonomic bootstrap of the clade sister to LECA. To gain further insights about the assigned donor, we checked its proportion in the first sister and its presence in the second sister.

To assess the signals that emerged through vertical evolution of the Asgard archaeal and Alphaproteobacterial donors we repeated the whole LECA prediction process as explained above. First we collected a set of representative 35 Asgard archaea and 35 Alphaproteobacterial proteomes, then we filtered the pangenome to exclude the species of interest so that each proteome set had its own pangenome minus sequences of its taxon. Tree reconstruction and identification of ancestral Asgard archaeal and Alphaproteobacterial groups were done as detailed above. An ancestral group was defined by the presence of five Asgard archaea or Alphaproteobacteria species depending on the dataset.

3.3. Inferring the acquisition waves

We define a LECA gene acquisition wave as a set of genes potentially acquired from the same donor during the same period of time. To identify such acquisition waves, we computed the normalised stem lengths as in[3] and plotted the distributions of these branch length values with respect to the potential prokaryotic donor. We obtained the acquisition modules for those non-eukaryotic groups that transferred at least 15 genes to the proto-eukaryote in any eTOLDB and under a seven supergroups criterion and with a support higher than 95% (9 partners, 6 bacterial clades, 2 archaeal phyla and a viral clade). We inferred the acquisition module using the density function, the genes that are around the peak of the distribution are selected as members of the acquisition module. To obtain these genes, we set the threshold to the first minimum after the first maximum or, in its absence, the first inflexion point of the density curve. Those genes that transferred before the threshold (are in the peak) are included in the module. In contrast, the ones having longer stem lengths are discarded, assuming that they were transferred posteriorly or the values result from heterotachy or ghost lineages[23,41] .

To infer the relative timing of the transfers and their timeline, we inferred the mode of the stem length distribution using a Bayesian framework[23]. In brief, this process uses a Markov Chain Monte Carlo algorithm to infer the posterior distribution for the parameters, mode and variance of a Gamma distribution, which we observed that properly fits the empirical

distribution of the stem lengths. We used a pairwise comparison of the posterior distributions of the mode from different donors (assumed as relative ages) to establish a probabilistic timeline for the different transfers to the proto-eukaryote.

3.4. Inferring the metabolism and features of the donors

To infer the ancestral features of the organisms that transferred genes to LECA, for a given donor group, we first selected the genomes sharing at least 50% of the genes transferred from that group to LECA in its acquisition wave, and annotated the genomes. These genomes constitute the "donor-like" genomes set. Second, we calculated the frequency of each KO present in these genomes. Then, we ran anvi'o v8[80] with each donor-like genome. For both the metabolic modules and the KOs, we calculated their prevalence, that is the proportion of genomes containing a given KO or module. In the case of the metabolic modules, we transformed the prevalence by multiplying it to the mean module completeness in all the donor's descendant genomes to take into account the completeness of the module. Using this metabolic reconstruction, we plotted the inferred donor's metabolism using ggKEGG v1.2.3[81]. We further annotated each protein of the donor-like genome using an HMM search against the COG database and related the results with the COG categories. Using the annotated COG categories, we calculated the over-representations of each COG category proportion for a given donor with respect to the LECA consensus proteome. In this case, the over-representation against LECA consensus proteome would mean the functional enrichment of the donor's contribution to LECA. Using the modules and the KOs prevalence, we assessed the presence and absence pattern of features relevant to the different eukaryogenesis models for all the donors. No statistical contrast for the enrichment was used as the sample sizes are non-homogeneous and, therefore, they do not provide enough statistical power, though we can study the general trends.

3.5. Inferring Eukaryotic Signature Proteins (ESPs)

To define a new set of Eukaryotic Signature Proteins, we selected the KOs that were marked as innovations in all three datasets, and that passed the five supergroup criterion in at least one of them. Additionally, we cross-checked that no other mLECA-OG annotated with the same KO was marked as an acquisition in any of the datasets.

Methods References

45. Brown, M. W. *et al.* Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biol. Evol.* **10**, 427–433 (2018).
46. Lax, G. *et al.* Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* **564**, 410–414 (2018).
47. Tikhonenkov, D. V. *et al.* Microbial predators form a new supergroup of eukaryotes. *Nature* **612**, 714–719 (2022).
48. Torruella, G., Galindo, L. J., Moreira, D. & López-García, P. Phylogenomics of neglected flagellated protists supports a revised eukaryotic tree of life. *bioRxiv* 2024.05.15.594285 (2024) doi:10.1101/2024.05.15.594285.
49. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of

Eukaryotes. *Trends Ecol. Evol.* **35**, (2020).
50. Kitts, P. A. *et al.* Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.* **44**, D73–80 (2016).
51. Richter, D. J. *et al.* EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. Preprint at https://doi.org/10.1101/2020.06.30.180687.
52. Gao, X. *et al.* The P10K database: a data portal for the protist 10 000 genomes project. *Nucleic Acids Res.* **52**, D747–D755 (2024).
53. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
54. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Res.* **52**, D891–D899 (2024).
55. Wong, E. D. *et al.* Saccharomyces genome database update: server architecture, pan-genome nomenclature, and external resources. *Genetics* **224**, (2023).
56. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
57. Shen, W., Sipos, B. & Zhao, L. SeqKit2: A Swiss army knife for sequence and alignment processing. *Imeta* **3**, e191 (2024).
58. Wootton, J. C. & Federhen, S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149–163 (1993).
59. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
60. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
61. Maistrenko, O. M. *et al.* Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *ISME J.* **14**, 1247–1259 (2020).
62. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
64. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
65. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
66. Fuentes, D. *et al.* PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res.* **50**, D1062–D1068 (2022).
67. Collingridge, P. W. & Kelly, S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics* **13**, 117 (2012).
68. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat. Commun.* **13**, 6968 (2022).
69. Lassmann, T. Kalign 3: multiple sequence alignment of large datasets. *Bioinformatics* **36**, 1928–1929 (2019).
70. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated

alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

71. 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).
72. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
73. Kosiol, C. & Goldman, N. Different versions of the Dayhoff rate matrix. *Mol. Biol. Evol.* **22**, 193–199 (2005).
74. Müller, T. & Vingron, M. Modeling amino acid replacement. *J. Comput. Biol.* **7**, 761–776 (2000).
75. Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* **18**, 691–699 (2001).
76. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
77. Le, S. Q., Dang, C. C. & Gascuel, O. Modeling protein evolution with several amino acid replacement matrices depending on site rates. *Mol. Biol. Evol.* **29**, 2921–2936 (2012).
78. 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).
79. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
80. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nat. Microbiol.* **6**, 3–6 (2021).
81. Sato, N., Uematsu, M., Fujimoto, K., Uematsu, S. & Imoto, S. ggkegg: analysis and visualization of KEGG data utilizing the grammar of graphics. *Bioinformatics* **39**, (2023).

Acknowledgements

The authors want to thank members of the Gabaldón lab for the useful discussions, particularly Giacomo Mutti for his help in the taxonomy of the eukaryotic groups and Eduard Ocaña-Pallarés for the list of publicly available genomes of osmotrophs. We also want to thank Eugene Koonin for his comments on the manuscript. This project was mainly funded thanks to the support of the Gordon and Betty Moore Foundation (grant number GBMF9742). TG group also acknowledges support from the Spanish Ministry of Science and Innovation (grant numbers PID2021-126067NB-I00, CPP2021-008552, PCI2022-135066-2, PLEC2023-010225, and PDC2022-133266-I00), cofounded by ERDF "A way of making Europe", as well as support from the Catalan Research Agency (AGAUR) (grant number SGR01551); ; "La Caixa" foundation (grant number LCF/PR/HR21/00737), Fundació La Marató de TV3 (202328-31), AECC (PRYGN234923GABA), and Instituto de Salud Carlos III (IMPACT grant IMP/00019 and CIBERINFEC CB21/13/00061- ISCIII-SGEFI/ERDF).

Authors contributions

MB, SM-M, and MM-H performed all bioinformatics analyses, wrote the necessary code, and produced the visualisations. TG conceptualised the project, obtained funding and supervised the project. All authors contributed to the design of the analytical steps, to the interpretation of the results, and to writing the manuscript.

Competing interest declaration

The authors declare no competing interest.

Data availability

All the alignment, trees, annotations, and supporting files to carry on this research are deposited in a Zenodo repository that will be available upon publication.

Code availability

A GitHub repository with the code to reproduce our analysis will be available upon publication.

Extended Data legends

Extended Data Figure 1. COG categories distribution for free living unicellular eukaryotic genomes with different trophic strategies. Each panel shows a trophic strategy, FLUA: autotrophs, FLUO: osmotrophs and FLUP: phagotrophs. Each point refers to a genome with a given trophic strategy and the black lollipop shows the proportion of that category in the consensus proteome of LECA.

Extended Data Figure 2. Probability of the waves occurring at different evolutionary relative time points. Each heatmap value shows the pairwise probability of the wave from the donor in the row happening before the wave of the donor in the column. The donors are sorted by the acquisition time from older to more recent.

Extended Data Figure 3. Non-eukaryotic acquisition waves in LECA. Posterior distribution of the mode of the stem length distribution for each putative donor. The panels show the results from different databases (in rows) and selection criteria (in columns).

Extended Data Figure 4. Prevalence of the pathways and enzymes in the genomes of the donor's descendants. Each box shows a group of metabolic pathways and enzymes. The opacity of the colour shows the prevalence (the percentage of genomes with the feature) of that specific pathway or enzyme in the genomes from the donors that share at least 50% of the KOs that they transferred to LECA.