Dominant contribution of Asgard archaea to eukaryogenesis

Victor Tobiasson[1,*], Jacob Luo[1], Yuri I Wolf[1], Eugene V Koonin[1,*]

1. Computational Biology Branch, Division of Intramural Research, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*Correspondence to: victor.tobiasson@nih.gov or koonin@ncbi.nlm.nih.gov

Abstract

Origin of eukaryotes is one of the key problems in evolutionary biology. The demonstration that the Last Eukaryotic Common Ancestor (LECA) already contained the mitochondrion, an endosymbiotic organelle derived from an alphaproteobacterium, and the discovery of Asgard archaea, the closest archaeal relatives of eukaryotes inform and constrain evolutionary scenarios of eukaryogenesis. We undertook a comprehensive analysis of the origins of the core eukaryotic genes tracing to the LECA within a rigorous statistical framework centered around evolutionary hypotheses testing using constrained phylogenetic trees. The results reveal dominant contributions of Asgard archaea to the origin of most of the conserved eukaryotic functional systems and pathways. A limited contribution from Alphaproteobacteria was identified, primarily relating to the energy transformation systems and Fe-S cluster biogenesis, whereas ancestry from other bacterial phyla was scattered across the eukaryotic functional landscape, without consistent trends. These findings suggest a model of eukaryogenesis in which key features of eukaryotic cell organization evolved in the Asgard ancestor, followed by the capture of the Alphaproteobacterial endosymbiont, and augmented by numerous but sporadic horizontal acquisition of genes from other bacteria both before and after endosymbiosis.

Introduction

The origin of eukaryotes (eukaryogenesis) is undoubtedly one of the central problems in evolutionary biology [1-5]. Eukaryotes drastically differ from archaea and bacteria (collectively, prokaryotes) due to the highly complex organization of the eukaryotic cells. The signature features of this organizational complexity include the endomembrane system, in particular, the nucleus, the eponymous eukaryotic organelle, the elaborate cytoskeleton, and the mitochondrion, the energy-converting organelle that evolved from an Alphaproteobacterial endosymbiont [6]. Early models of eukaryogenesis featured a protoeukaryote lacking mitochondria that captured and domesticated an alphaproteobacterium at a relatively late stage of evolution [7,8]. However, subsequent research revealed no primary amitochondrial eukaryotes (although many have lost mitochondria through reductive evolution), suggesting that the Last Eukaryotic Common Ancestor (LECA) was already endowed with mitochondria along with the other signatures of the eukaryotic cellular organization [1,3]. These findings lend credence to scenarios of eukaryogenesis in which mitochondrial endosymbiosis triggered cellular reorganization, giving rise to the eukaryotic cellular complexity [3,9-11].

Phylogenomic analyses show that those conserved, core genes of eukaryotes that can be traced to the LECA are a mix originating from archaea and from various bacteria. The archaea-derived genes were found primarily in information processing systems (replication, core transcription and translation) whereas genes of apparent bacterial origin comprise the operational component of the eukaryotic gene complement, in particular, encoding metabolic enzymes [12-15].

The concepts of eukaryogenesis was transformed by the discovery and subsequent exploration of Asgard archaea (currently phylum Promethearchaeota within kingdom Promethearchaeati [16] hereafter, Asgard, for brevity) that includes the closest known archaeal relatives of eukaryotes [17-21]. Phylogenetic analyses of highly conserved genes have been pushing the eukaryotic branch progressively deeper with the Asgard tree. The latest such analysis identified the order Hodarchaeales, within the Asgard class Heimdallarchaeia, as the likely sister group of eukaryotes [21]. Perhaps, even more notably, it has been shown that Asgard archaea encode homologs of a broad variety of eukaryote signature proteins beyond the core information processing componentry, in particular, many cytoskeletal proteins and those involved in membrane remodeling [20-22].

A variety of eukaryogenesis models have been proposed, differing with respect to the timing of the origin of the typical eukaryotic cellular organization and the exact nature and succession of the underlying events [1,3]. The endosymbiotic origin of the mitochondria from an alphaproteobacterium is indisputable, but the nature of the host of the proto-mitochondrial endosymbiont remains a matter of debate. The most straightforward models posit an Asgard archaeal host. However, a major conundrum for such scenarios of eukaryogenesis is the chemistry of cell membrane lipids and the enzymology of their biosynthesis, which are unrelated in archaea and bacteria, with eukaryotic being of the bacterial type [23]. Accordingly, any model of eukaryogenesis that includes an archaeal host would require a membrane replacement step.

More complex, alternative models of eukaryogenesis, underpinned by metabolic symbiosis, postulate two endosymbiotic events whereby an Asgard archaeon was first engulfed by a bacterium, followed by the loss of the archaeal membrane, and then, a second endosymbiosis rise to the mitochondria [24-26]. Even if seemingly less parsimonious than single-endosymbiosis scenarios, such models of eukaryogenesis account for the continuity of bacterial membranes and also appear compatible with the syntrophic lifestyle of at least some Asgard archaea, which have been found in consortia with bacteria, Myxococcota (formerly Deltaproteobacteria), in particular [27,28]. Furthermore, a pivotal study by Pittis and Gabaldon in which relative timing of events in eukaryogenesis was inferred by comparing stem length from different ancestors in phylogenetic trees of conserved eukaryotic genes, suggested a late capture of the mitochondria, apparently preceded by acquisition of many genes from different bacteria, possibly, through an earlier endosymbiosis [29,30].

Here, we took advantage of the rapidly growing collection of archaeal, bacterial and eukaryotic genome sequences to infer the origins of core eukaryotic genes tracing to the LECA within a rigorous statistical framework based on evolutionary hypotheses testing using constrained phylogenetic trees. The results reveal the dominant contribution of Asgards to the origin of most functional classes of eukaryotic genes, which is compatible with a scenario of eukaryogenesis where many signature, complex features of eukaryotic cells evolved in the Asgard ancestor of eukaryotes prior to mitochondrial endosymbiosis.

Main

Phylogenomic analysis and global evolutionary associations of conserved eukaryotic proteins

To identify associations between prokaryotic and eukaryotic protein families, separate hidden Markov model (HMM) databases for prokaryotes and eukaryotes were constructed using a custom, cascaded, sequence-to-profile clustering pipeline, implemented using mmseqs2 [31]; followed by a multistep data-reduction and multiple sequence alignment (MSA) procedure to generate HMM profiles using hhsuite [32] (see Methods for details). Initially, a prokaryotic database of 37 million protein sequences was curated from 24,757 complete prokaryotic genomes obtained from the NCBI GenBank in November 2021 and supplemented with proteins extracted from 146 Asgard genome assemblies (supplementary Figure 1) [19,20]. To avoid inclusion of genes present only within a narrow subset of species, possibly resulting from horizontal transfer from eukaryotes post LECA, we reconstructed the "soft-core" pangenome for each of the 26 curated prokaryotic taxonomic classes. These pangenomes include only those genes that are present in at least 67% of the families within each class of bacteria and archaea (see Methods) resulting in an initial database of 13 million sequences. The initial eukaryotic database consisted of 30 million protein sequences from 993 species taken from EukProt v3 [33] and cleaned using mmseqs2 to remove likely prokaryotic contaminants, followed by clustering and MSA construction (see methods).

The resulting eukaryotic HMM dataset was queried against the prokaryotic dataset using hhblits to identify sets of homologous protein sequences.
Each eukaryotic cluster and all its significant prokaryotic hits constituted an individual sequence set, hereinafter referred to as a Eukaryotic/Prokaryotic Orthologous Cluster (EPOC). The EPOCs constitute groups of homologous proteins from eukaryotes and prokaryotes (each EPOC contains a unique set of eukaryotic proteins, but some clusters of prokaryotic proteins can be present in multiple EPOCs) that were used for phylogenetic tree construction, annotation, and evolutionary hypothesis testing. The final EPOCs include 10.4 million prokaryotic and 2.7 million eukaryotic sequences, mapping to 80 % and 9 % of the datasets, respectively.

To infer the most likely prokaryotic ancestry of the eukaryotic proteins in each EPOC, rather than relying on the tree topology directly, we employed a probabilistic approach for evolutionary hypothesis testing using constraint trees. We exhaustively sampled all arrangements of likely sister clades and obtained Expected Likelihood Weights (ELW) for the set of possible sister clade models, (supplementary Figure 2) [34]. As the ELW metric is analogous to model selection confidence, here we take it to be proportional to the probability of a sampled prokaryotic clade to be the true sister group of the given eukaryotic clade among a set of competing sister clades. For each EPOC, our analysis dynamically accounts for long branch outliers and is robust to phylogenetically non-homogenous clades (see methods). This analysis is further capable of resolving eukaryotic paraphyly, treating each eukaryotic clade within a EPOC as a single datapoint for downstream analysis. Our resulting data included 18,800 EPOCs annotated using profiles generated from KEGG Orthology Groups (KOGs) [35], each with an MSA generated using muscle5 [36], a maximum likelihood tree inferred using IQtree2 [37] and associated ELW values for all candidate prokaryotic sister phyla. The analysis of prokaryotic ancestry was performed only for those eukaryotic clades that included more than 5 distinct taxonomic labels, with at least one coming from Amorphea and one from Diaphoretickes, the two expansive eukaryotic clades considered to represent either the first or the second bifurcation in the evolution of eukaryotes [38,39]. Thus, these clades likely represent genes mapping back to the LECA.

Considering the global average distribution of ELW values across all EPOCs covering 5060 unique KOGs, the single greatest average ELW (aELW), here referred to as an association, was with Asgard archaea. Further associations with Cyanobacteria, Actinomycetota and Alphaproteobacteria, as well as trace association with diverse bacteria, were detected at lower levels. However, the number of included sequences and topological diversity of the trees varied substantially across the EPOCs, with many trees showing low maximum ELW values. Excluding EPOCs with a maximum ELW < 0.4 yielded a robust core set of 6,990 EPOCs stemming from the LECA, covering 3050 KOGs and improving the interpretability of the results. This core set covers a wide range of ubiquitous metabolic functions, information processing pathways, transporters, and permeases, as well as regulatory and housekeeping proteins.

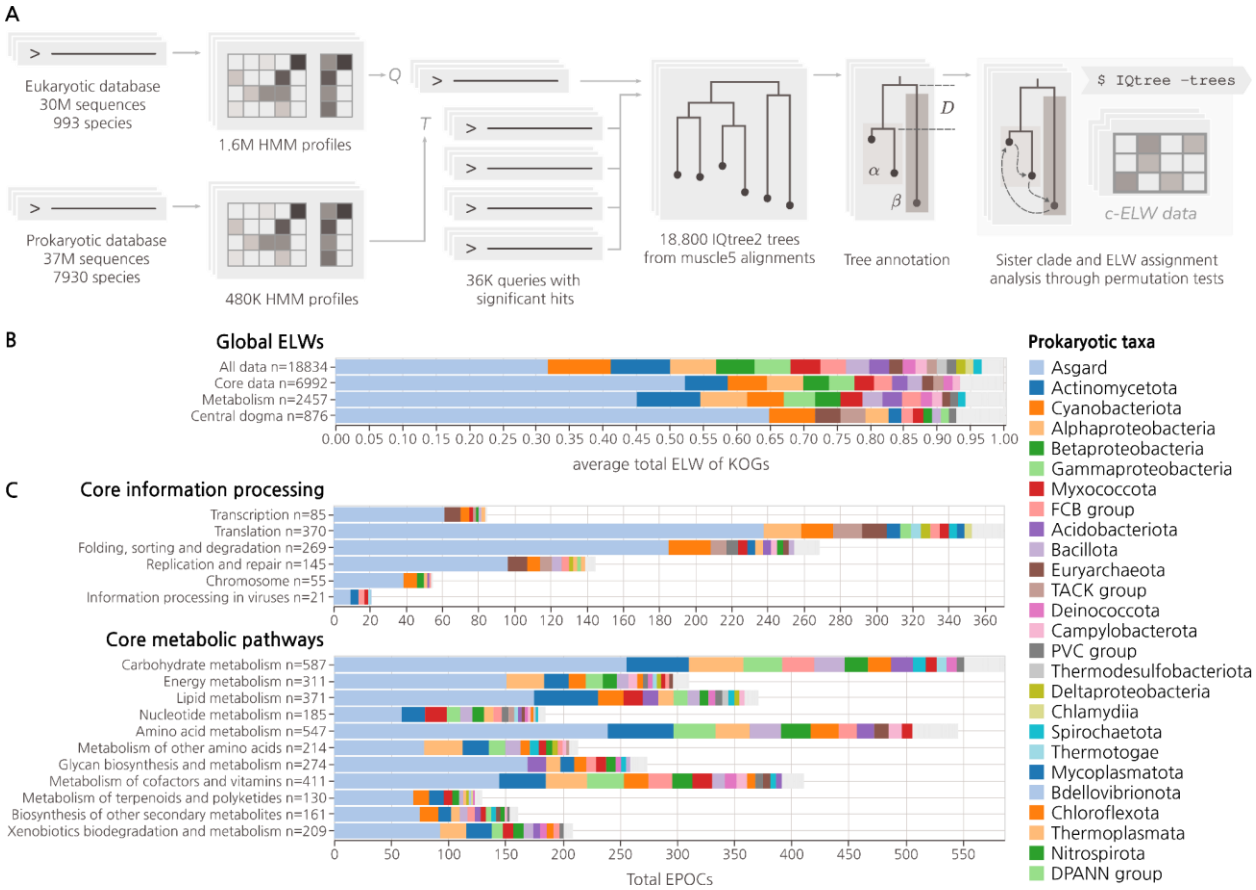


Figure 1. Phylogenetic associations of functional classes of eukaryotic proteins with major archaeal and bacterial taxa. A) Simplified pipeline consisting of, from left to right, one eukaryotic and one prokaryotic sequence database transformed into HMM profiles. The query of eukaryotic against prokaryotic profiles generates 18.800 valid EPOCs, automatically annotated and with detected prokaryotic clades assessed for their probability to be the closest eukaryotic sister phyla by permutation testing using IQTree2. **B)** Global average ELW for all EPOCs within the unfiltered and core datasets as well as top level breakdowns for metabolic enzymes and proteins involved in the central dogma as defined by the KEGG ontology "BRITE" at the top "A" level (KO:09100, KO:09120). **C)** Further subdivision of average ELW values across the lower BRITE "B" level for each BRITE A category.

By contrast, it does not include metabolic pathways limited to individual eukaryotic supergroups and thus can be considered a rough approximation of the LECA gene set. Limiting the analysis to this core subset of well assigned eukaryotic families with wide taxonomic coverage notably increased the global Asgard association which now accounted for the 52 % of all likelihood weight across almost 7000 unique observations in more than 3000 protein families.

Our approach readily reproduced known evolutionary associations at the global functional level. Averaging ELW scores across EPOCs based on the KEGG ontology shows support for major Asgard association for, among others functional systems and pathways: the ribosome, RNA and DNA polymerases, Ras-like GTPases, Ubiquitin mediated protein degradation, the proteasome, and large parts of the core metabolic network. In contrast, prominent Alphaproteobacterial associations included proteins involved in oxidative phosphorylation, glutathione metabolism and Fe-S clusters biogenesis. Together, the Asgard and Alphaproteobacterial associations amounted to an aELW of 0.56 + 0.05 = 0.61 across all of Metabolism (KO:09100) and to 0.50 + 0.06 = 0.56 across Genetic Information Processing (KO:09120). Thus, we observed a far stronger overall association between Asgards and eukaryotes across diverse biological functions and pathways than previously described [18,20,21] although associations between eukaryotic core metabolism and diverse bacterial phyla were still present.

Broad, dominant Asgard contributions to eukaryogenesis

In accord with the core contribution of Asgard archaea to eukaryogenesis, we observed associations of Asgard proteins with a wide array of cellular functions. In previous work, the strongest Asgards traces have been noted across the information processing systems, with unambiguous protein associations with DNA and RNA processing as well as protein expression, trafficking and signaling [2,20,21]. Here, we consistently observed strong Asgard associations for genome replication and transcription and further detect pronounced Asgard traces for nucleotide excision repair, mismatch repair and homologous recombination. Additional well-known associations such as Ribosomal proteins are extended to include translation factors, components of the co-translational insertion machinery, protein targeting and Aminoacyl-tRNA biosynthesis (supplementary Figure 3, supplementary data Y). Thus, all groups of core eukaryotic proteins involved in information processing appear to be almost exclusively of Asgard descent.

We detected further Asgards associations extending far beyond the information processing systems, including prominent contributions of Asgard archaea to the machinery involved in nucleocytoplasmic transport as well as downstream protein sorting, glycosylation and targeting. In particular, central components of the ER associated, N-linked glycan biosynthesis and transfer, including both cytoplasmic and lumenal monoglycosyltransferases, as well as the core of the oligosaccharyltransferase complex (OSTC) are strongly associated with Asgards (supplementary Figure 3). Notably, enzymes associated with glycosylation maturation in the Golgi complex did not show strong Asgard or other prokaryotic associations in our analysis, indicating absence of readily detectable prokaryotic ancestors, possibly, due to evolution of extensive domain diversity in these enzymes.

202 The Asgard connections to glycosylation further included the synthesis of GPI-anchors, which
203 post-translationally tether targeted proteins to the membrane, here detected as unambiguously
204 Asgard-derived. We also detected an Asgard origin of the 7-subunit (UDP-GlcNAc)-transferring
205 (GPI-GnT)-monoglycosyltransferase complex responsible for initiating GPI-anchor synthesis,
206 components required for the maturation of the GPI-anchor, as well as the transamidase complex
207 and factors responsible for transfer of the protein onto the mature GPI-anchor (supplementary
208 Figure 3).
209
210 Of major importance to eukaryogenesis is the provenance of the pathways for the biosynthesis
211 of bacterial-type lipids, and all binary archaea-bacteria symbiogenesis scenarios require a
212 transition from archaeal to bacterial lipids in the membranes. Although strong Asgard associations
213 were observed for large parts of the overall metabolic network, we observed a high degree of
214 mosaicity in the pathways for fatty acid synthesis and decay. The global aELW values favored
215 Asgard origin for these pathways, but there were also notable associations with Actinomycetota,
216 and most KOGs within these pathways are represented by multiple EPOCs with conflicting
217 assignments, obscuring any constructive signal, (supplementary Figure 3). However, less
218 perplexity was observed within the adjoining ER localized pathways for sphingolipid metabolism,
219 a broad class of derived plasma membrane lipids in eukaryotes. Previous studies have highlighted
220 possible convergent origins of this pathway in bacteria and eukaryotes, but here we detected
221 broad associations with Asgard archaea. Of further note is the ER-associated isoprenoid
222 biosynthesis pathway, here also found to be strongly associated with Asgard. In eukaryotes,
223 isoprenoids form the precursor units for both sterols, carotenoids and terpenoids, synthesized in
224 the ER lumen via either the mevalonate pathway or the MEP/DOXP pathway. In Archaea,
225 isoprenoids are the precursors for the ether-linked membrane lipids. Here we found the
226 mevalonate pathway, from Acetyl-COA to mevalonate and further to Farnesyl and Geranyl
227 diphosphate, to be strongly Asgard-associated (supplementary Figure 3) with all of the detected
228 enzymes closely associated to Asgards (HMGCS, MVK, PMVK, MVD).
229
230 In conclusion, we detected Asgard associations across a wide range of cellular functions and
231 metabolic pathways while noting a distinctly weaker Asgard signal for pathways involved in
232 bacterial lipid biosynthesis, suggesting a complex evolutionary history.
233
234

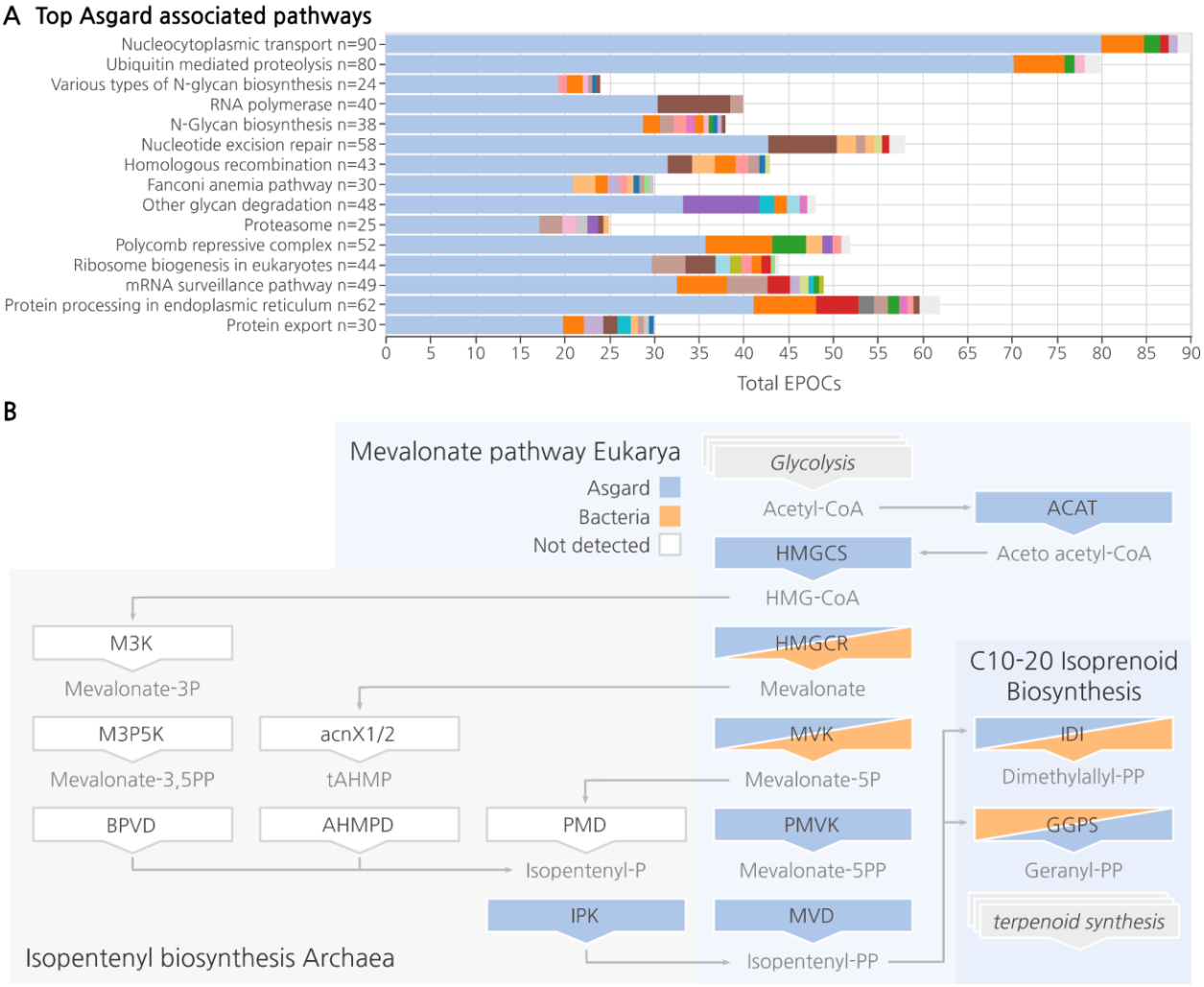
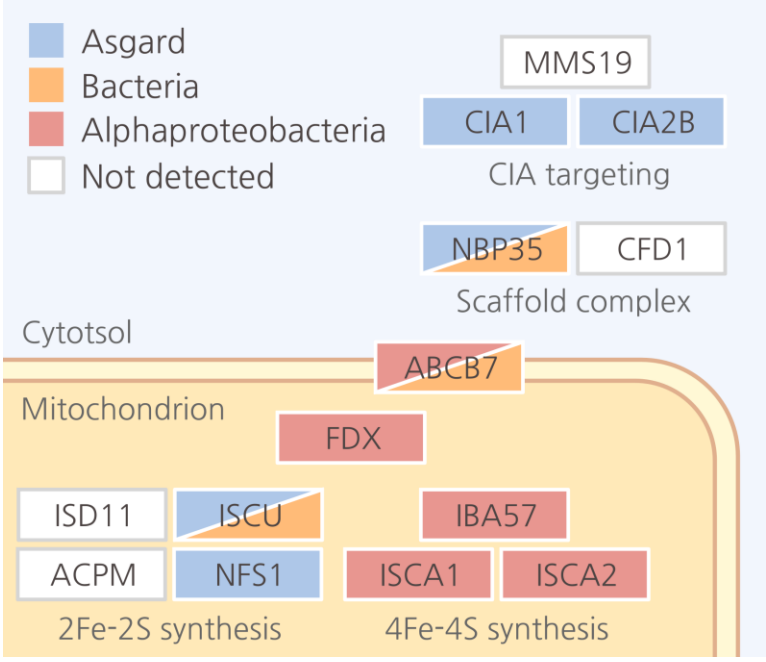


Figure 2. Asgard origins of eukaryotic pathways. A) Top 15 KEGG maps with strongest Asgard association, only concludes maps with more than 20 EPOCs. **B)** DEMO. Mevalonate pathway in eukaryotes with each protein colored by inferred ancestry. Split backgrounds correspond to mixed signals.

Limited and highly specific Alphaproteobacterial contributions

In line with the central role of mitochondria in eukaryotic energy metabolism, we primarily observed associations between Alphaproteobacteria and mitochondrially localized metabolic pathways. As expected, apart from the components of the mitochondrial translation system, the most prominent Alphaproteobacterial associations were evident within complexes involved in oxidative phosphorylation and the associated ubiquinone synthesis (supplementary Figure 3). Outside these central energy-transforming functions, we only detected sparse contributions from core Alphaproteobacterial genes. One such prominent association was the pathway for iron sulfur cluster (ISC) biogenesis. As previously noted [40], the ISC assembly machinery is alphaproteobacterial origin, and in accord with these observations, the 4Fe-4S platforms ISCA1, ISCA2 and IBA57, as well as Fe-S cluster binding ferredoxin-1 and 2, were found to be strongly associated with Alphaproteobacteria (Figure 3, supplementary Figure 3). However, the 2Fe-2S

253 precursor scaffold ISCU showed mosaic associations, with a minor but clearly detectable Asgard
254 contribution. Notably, the cysteine desulfurase NFS1 and the upstream pathways for the
255 biosynthesis of sulfur-containing amino acids, cysteine, and methionine, were strongly Asgard
256 associated. The ISC biosynthesis is intimately associated with general redox homeostasis and
257 core sulfur metabolism via glutathione, directly coordinating Fe-S clusters during synthesis and
258 transport. In line with this role in Fe-S coordination, although some Asgard association persisted,
259 we observed specific associations of glutathione metabolism with Alphaproteobacteria, including
260 both glutathione hydrolase, dehydrogenase and reductase, and the family of glutathione
261 transferases, GST, GSTP and GSTK1 (supplementary Figure 3). Outside the mitochondria, ISC
262 insertion depends on the cytosolic targeting complex CIA, which consists of CIAO1, CIA2B and
263 MMS19 [41]. For the CIA components CIA1 and CIA2B, we observed clear association with Asgard
264 whereas the origin of MMS19 remains unresolved. Taken together, these observations indicate
265 that the contributions of Alphaproteobacteria to the gene set of the LECA are functionally specific
266 and likely limited in scope, yet clearly centered around mitochondria-related functions.
267



268
269 **Figure 3. Ancestry of Eukaryotic Fe-S synthesis pathways.** Schematic for Fe-S cluster
270 assembly including downstream cytosolic insertion. Specific ancestry of the mitochondrial 4Fe-4S
271 synthesis and assembly pathways from Alphaproteobacteria. Cytosolic or cytosolic-derived
272 components associated with Asgards.
273

Absence of functionally consistent contributions from other bacteria

Although our analysis greatly expands the Asgard contributions to eukaryogenesis, while also revealing a limited but prominent and functionally consistent Alphaproteobacterial association, contributions from diverse other bacteria were detected. For some biological functions, this diverse bacterial component accounted for the majority of the aELW, and roughly one third of the analyzed KOGs (918 of 3045), and EPOCs (2640 of 6990) were associated neither with known ancestors of endosymbionts, Alphaproteobacteria and Cyanobacteria, nor with Asgard. However, in a sharp contrast to Archaeal associations including information processing, protein glycosylation and trafficking, and others, or oxidative phosphorylation and sulfur metabolism for Alphaproteobacteria, EPOCs associated with diverse other bacteria showed few if any coherent functional trends.

Considering the diverse set of bacteria, only Actinomycetota, Alphaproteobacteria and Cyanobacteria were associated with at least one pathway containing more than 20 EPOCs and with a greater aELW than Asgards (figure 4). Among prominent Actinomycetota associations, 7 of the top 10 were pathways involved in fatty acid or lipid biosynthesis, including arachidonic acid metabolism, and glycerophospholipid biosynthesis. However, Actinomycetota had the highest aELW in only one of these cases (arachidonic acid), closely followed by Alphaproteobacteria. For all other pathways, Asgard associations dominated. Additionally, the specific enzymes detected poorly covered the respective pathways (6 KOGs out of 51 possible for Arachidonic acid metabolism) and specific contributions remained scattered. The second most prominent bacterial contributor is Gammaproteobacteria, with associations scattered across multiple pathways including, among others, lipid, amino acid, sulfur, nicotinate, and pyruvate metabolism. However, in no case did we detect the Gammaproteobacterial clade as the strongest association, which was typically Asgard, and in all cases, other competing bacterial clades were present at comparable levels (supplementary Figure 4).

Most bacterial phyla were associated primarily with diverse metabolic pathways, without clearly observable trends, but several prominent exceptions were notable. Although still weaker than Asgards, Myxococcota was most consistently associated with nicotinate, nucleotide, and nucleoside sugar metabolism. However, these associations were limited to phosphatases and phosphoribosyltransferases acting on nucleotide sugars including 5 and 3′ nucleotidases, with the relevant EPOCs showing little to no competing Asgard associations. While noteworthy, these associations were limited to a few unique KOGs whereas all associations with Myxococcota remained scattered across various pathways (supplementary Figure 5).

In addition to investigating metabolic pathways for global associations, we directly examined those individual EPOCs that were highly likely to be derived from diverse bacteria. For this analysis, we considered a stricter subset of the core EPOCs, requiring a eukaryotic outgroup containing at least 15 taxonomic clades and prokaryotic sister taxa with at least 20 sequences and ELW > 0.7.

Only 30 of the 228 unique KOGs meeting these criteria were found to be associated with diverse bacterial lineages, mostly, Actinomycetota, FCB group and Betaproteobacteria, with minor contributions from Mycoplasmatota and Campylobacteriota (supplementary Figure 6). The remaining 198 were Asgard derived. These prominent bacterial KOGs covered a wide range of cellular functions from MFS transporters to lipases to components of core sugar metabolism and cardiolipin synthesis once again with no noticeable trends. Taken together, although many associations were individually significant, no pathways appeared significantly enriched in associations with any single bacterial taxon, and we identified no consistent trends. Instead, we interpret these systemic but diffuse associations to diverse prokaryotes to be suggestive of highly specific prokaryotic contributions of limited functional scope from outside of known endosymbiotic partners.

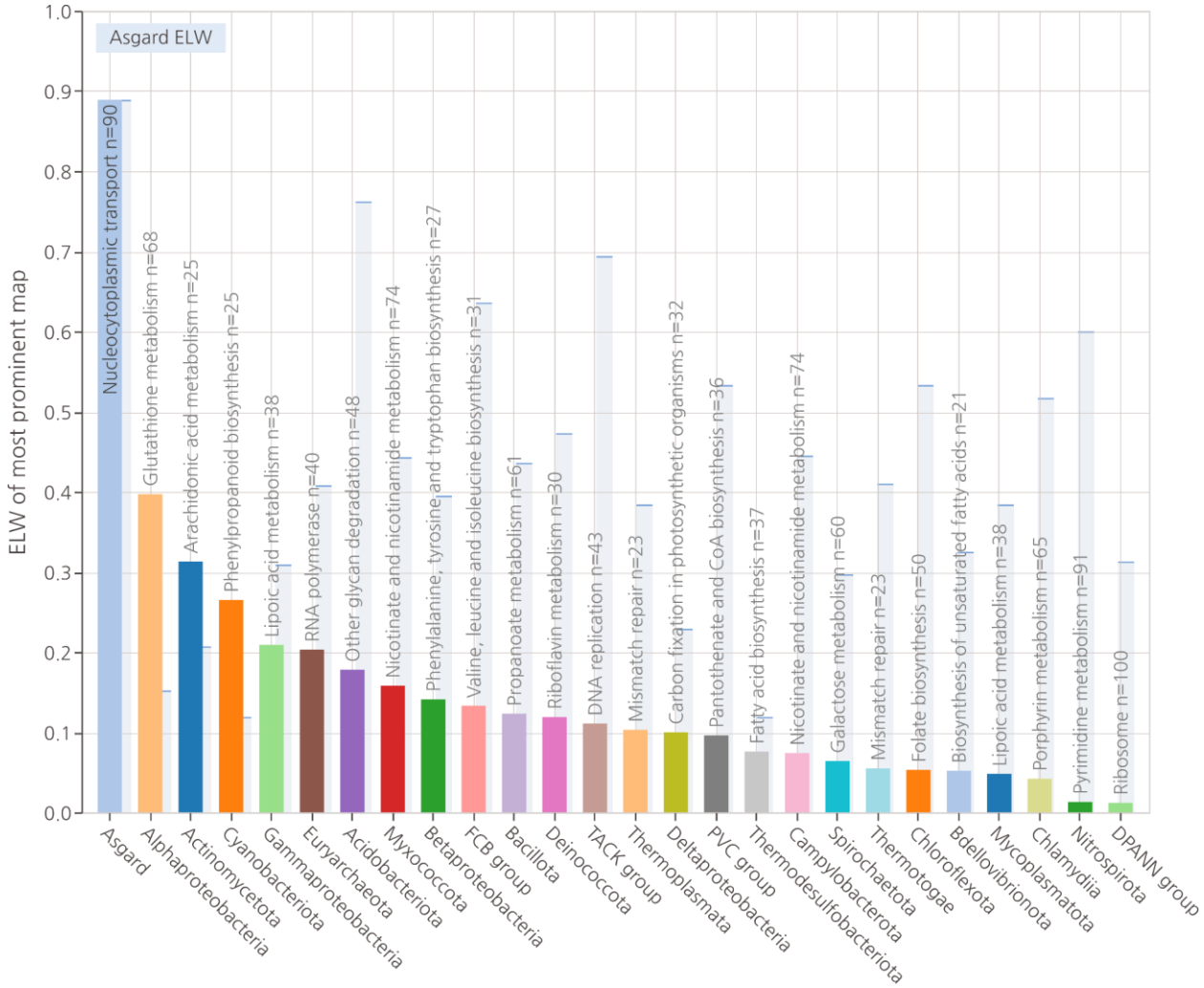


Figure 4. Diverse bacterial contribution to the origins of eukaryotic genes compared to the contributions of Asgard archaea. Maximum ELW for the single highest enriched KEGG maps per prokaryotic taxa, relative to Asgard contributions for each respective map, is shown.

334 **Relative amounts of evolution pre- and post-LECA**
335 We employed the methodology pioneered by Pittis and Gabaldon [30] to compare the ancestral
336 stem length in phylogenetic trees of core eukaryotic genes of different inferred origins. Briefly,
337 we define the raw stem length as the distance from the LECA node to the shared LCA node of
338 Eukarya and their most likely prokaryotic sister phyla. To account for differences in evolutionary
339 rates, we further divided this raw stem length by the median of eukaryotic branch lengths,
340 measured from the LECA to each leaf (Figure 5C). Considering that the astronomic time post-
341 LECA is the same for all genes, if the tempo and mode of pre- and post-LECA evolution are the
342 same, the normalized stem length is proportional to the time elapsed since the divergence of the
343 gene from its prokaryotic donor to LECA, that is, reflects the timing of acquisition. Using this
344 approach, Pittis and Gabaldon found that proteins of Alphaproteobacterial descent had
345 significantly shorter normalized stems than proteins of archaeal descent, consistent with a
346 mitochondria-late scenario for eukaryogenesis [30].
347
348 Across the full set of eukaryotic stem lengths for our dataset (5,956 stems), we observed a wide
349 distribution with a sharp maximum close to 0.1, highly reminiscent of the previous findings (Figure
350 5B). However, in our analysis, the stems for genes of Alphaproteobacterial origin were
351 significantly ($p \approx 3x10^{-7}$, see figure 5) longer than those of Asgard origin, and longer than those
352 of other major bacterial contributors as well (Figure 5B). Comparison of the stem length
353 distributions across functional classes of genes (Figure 5A) suggested an explanation for these
354 observations. The shortest stems belong to the Genetic Information Processing functional
355 category, belying our expectations of these genes being the longest-residing genes in the nascent
356 eukaryotic lineage. We suggest that another major factor, determining the relative stem length
357 of a gene, is the amount of adaptive evolution post-acquisition, which is necessary to adjust the
358 newly acquired gene to the alien intracellular molecular environment. Thus, genes inherited from
359 the Asgard ancestor, in particular, those involved in information processing, were pre-adapted to
360 the cellular environment of the evolving protoeukaryotes, whereas genes acquired from radically
361 different bacterial sources had to substantially adapt post-acquisition, increasing the apparent
362 lengths of their pre-LECA stems. A case in point is the set of OXPHOS components, which were
363 apparently acquired during the mitochondrial symbiogenesis, and thus, simultaneously from the
364 same donor. The distribution of their stem lengths (Figure 5A) was as wide as that for other
365 components of cellular machinery (Figure 5A and supplementary Figure 7), demonstrating that
366 stem lengths are mostly determined by factors other than the acquisition timing. Thus, our stem
367 length analysis failed to provide unequivocal resolution of the temporal order of the prokaryotic
368 contributions to eukaryogenesis. Nevertheless, the results appear to be compatible with the
369 capture of the Alphaproteobacterial endosymbiont by a host that was already on the path to
370 eukaryotic-like complexification, along with capture of genes from various bacteria at different
371 stages of eukaryogenesis.
372

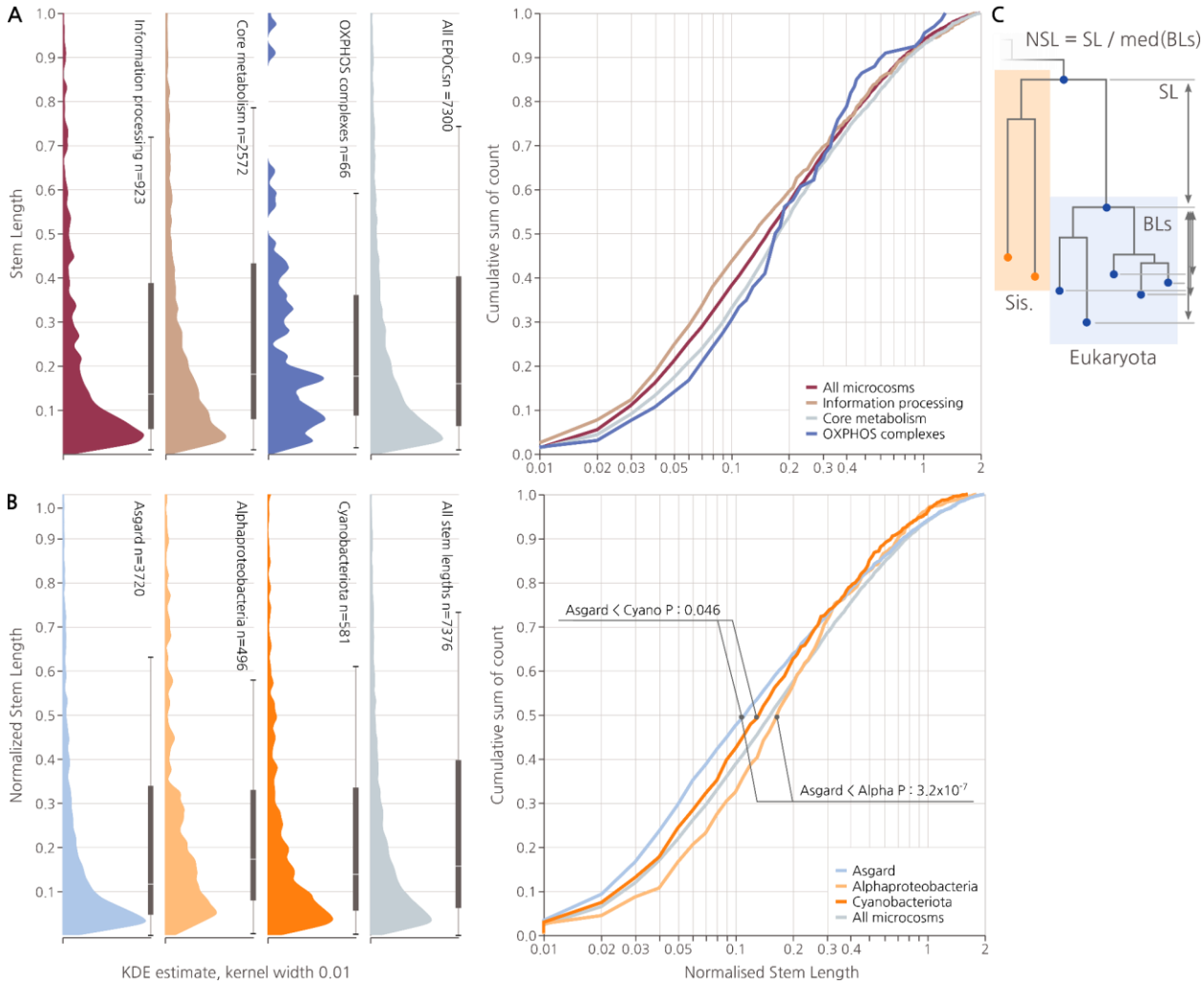


Figure 5. Ancestral stem length distributions for different groups of prokaryotes contributing to eukaryogenesis and different functional classes of genes. A) Distribution of normalized eukaryotic stem lengths for different top-level eukaryotic functions as defined by BRITE A Metabolism (KO:09100), Genetic Information Processing (KO:09120), Oxidative phosphorylation (map00190 excluding KOGs for V-type ATPases). Includes all eukaryotic clade stem lengths between 0.05 and 1.0, 79 % of total stems. **B)** Distribution for individual taxa with associations with eukaryotes, Actinomycetota showing enriched long stems compared to known contributors. Cyanobacteria and Alphaproteobacteria are observed with slightly shorter stems. E-values calculated using Mann-Whitney U test to assess whether the distribution underlying Asgards is stochastically less than the distribution underlying the indicated taxa. **C)** Relative amount of pre- and post-LECA evolution.

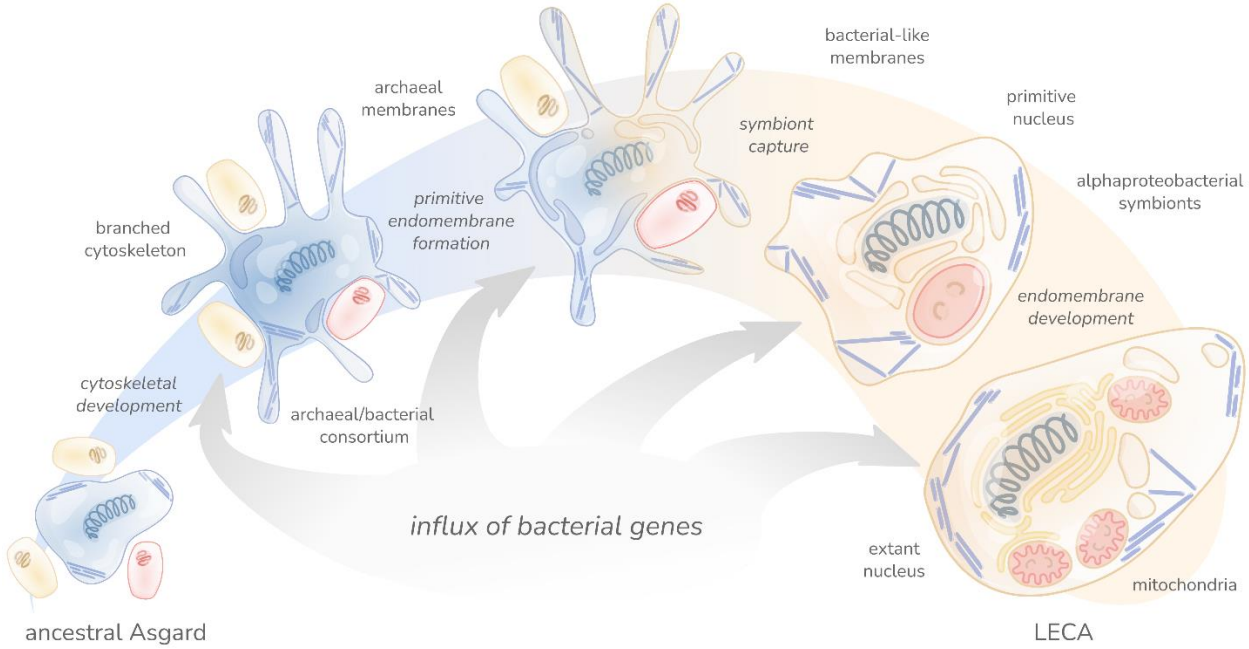
Discussion

Reconstruction of eukaryogenesis is a moving target as strikingly demonstrated by the discovery and ongoing exploration of Asgard archaea which revolutionized the field [17-21]. Nevertheless, recent progress in prokaryote as well as eukaryote genome sequencing, along with the advances of metagenomics, created unprecedented opportunities for evolutionary genomics study aimed at understanding the origins of eukaryotes. In this work, we took advantage of this expanded genome collection, coupling it with maximally sensitive HMM profile-profile searches, and performed a comprehensive phylogenetic analysis of core eukaryotic genes tentatively mapped to the LECA, within a rigorous statistical framework focused on testing evolutionary hypotheses using constraint trees.

The principal outcome of this analysis is that a substantial majority of the LECA genes, for which the origin could be inferred with confidence, came from Asgard archaea. Particularly notable is the functional diversity of the apparent Asgard contribution to eukaryogenesis. Far from being limited to information processing and certain cellular processes, such as membrane remodeling, a strong Asgard trace is detectable in most core functional systems and biochemical pathways of eukaryotes. In very few protein families did the detected bacterial contribution exceed that of Asgard archaea. The most conspicuous of the bacterial contributions, not unexpectedly, comes from Alphaproteobacteria, the ancestors of the mitochondria, and accounts for the core components of the electron transport chain complexes and iron-sulfur cluster metabolism, along with the mitochondrial translation system. However, although functionally relevant and consistent, the Alphaproteobacterial contribution is relatively small and not comparable in scale to that of Asgard. The contributions of other bacterial phyla, although cumulatively substantial, fail to show any individually consistent trends, with the exception of one or two scattered metabolic pathways.

Many previous studies, especially early ones, demonstrated the apparent chimeric origins of the core eukaryotic gene set, with the bacterial contribution quantitatively exceeding the archaeal one, and the latter being largely limited to information processing [13-15,42]. The discovery and exploration of Asgard archaea partially changed this notion by demonstrating that many genes involved in various cellular processes and systems, such as membrane remodeling and cytoskeleton, were of apparent Asgard origin [20-22]. The present work substantially extends the dominance of Asgard in the ancestry of the LECA genes, by revealing major Asgard contributions to nearly all functional systems and pathways of the eukaryotic cell. These findings imply a model of eukaryogenesis in which the ancestral Asgard archaeon already possessed many features characteristic of eukaryotic cells including cytoskeleton and the endomembrane system. These complex ancestral Asgard archaea might have led a predatory lifestyle conducive to engulfment of bacteria and serial capture of bacterial genes. Under this scenario, mitochondrial endosymbiosis, also perhaps facilitated by predation, while antedating the LECA, was a relatively late event that made a limited, even if functionally crucial, contribution to the gene composition of the emerging eukaryote (Figure 6).

429 Other bacterial contributions appear to be piecemeal, coming from a limited capture of genes
430 from diverse bacteria, likely, both before and after the mitochondrial endosymbiosis; there is no
431 indication of another symbiotic event contributing to eukaryogenesis. This conclusion
432 complements and expands a recent study that suggested emergence of some features of
433 eukaryotic complexity prior to the mitochondrial endosymbiosis based on timing of ancient gene
434 duplications [43].
435
436 However, how eukaryotes acquired the bacterial-type membranes, remains an enigma. Whereas
437 the eukaryotic pathways for isoprenoid synthesis are undoubtedly Asgard-derived, the ancestry
438 of the enzymes involved in membrane lipid biosynthesis appears complex and remains unresolved
439 in the present analysis, with multiple bacterial ancestors appearing likely. Our analysis of these
440 pathways is likely confounded by the presence of two sets of lipid biosynthetic pathways, one
441 mitochondrial and the other ER-associated. Nevertheless, the Asgard affinity of adjoining systems
442 and pathways, such as the synthesis of GPI anchors, sphingolipids, or isoprenoids, shows that a
443 substantial component of the lipid biosynthesis and modification capacity of eukaryotes persisted
444 from the Asgard ancestor.
445
446 Under the premise that an Asgard archaeal cell was the principal template on which the eukaryotic
447 cell evolved, replacement of the archaeal membrane with a bacterial one must have occurred at
448 an early stage of eukaryogenesis, likely, through a mixed membrane stage. Viable bacteria with
449 a mixed bacterial-archaeal membrane have been experimentally engineered albeit resulting in a
450 substantial fitness decrease [44,45]. Membrane replacement might have even predated mitochondrial
451 endosymbiosis, following early capture of the necessary bacterial enzymes by the Asgard
452 archaeon (Figure 6).
453



454
455 **Figure 6. Hypothetical scenario of eukaryogenesis**

The conclusions of this work are subject to caveats stemming, above all, from the biased and still limited sampling of sequenced archaeal and bacterial genomes. In particular, almost all currently available genomes of Asgard archaea are incomplete metagenomic assemblies, with the exception of only three closed circular genomes [28,46,47]. Therefore, even though our current analysis points to a dominant contribution of Asgard to the eukaryotic gene core, it is nevertheless almost certainly an underestimate. The same pertains to Alphaproteobacteria both because of the sequencing bias whereby many of the sequenced Alphaproteobacterial genomes come from symbionts and parasites with reduced gene complements, and because the ancestor of the mitochondria apparently belongs outside of the currently known diversity of Alphaproteobacteria [48]. Some of the other bacterial phyla are even more severely under-sampled, leading to underestimates of their contributions to the evolution of eukaryotes. For example, it is difficult to rule out that we miss a major signal from Myxococcota (formerly Deltaproteobacteria), that have been proposed as one of the partners in the syntrophy scenario of eukaryogenesis [24,25], given the limited genomic information on these bacterial phyla. The expanding sampling of prokaryotic genome diversity might lead to a substantial modification or even complete revision of the model of eukaryogenesis proposed here. It nevertheless appears likely that the dominant contribution of Asgard archaea, the main finding of this work, is here to stay.

Methods

Database curation

For prokaryotes ~37.0 million sequences were curated from 24,757 completely sequenced prokaryotic genomes obtained from the NCBI GenBank (https://ftp.ncbi.nlm.nih.gov/genomes/) in November 2021 and supplemented with 441,150 proteins sequences from 142 assembled Asgard genomes [20,21], selected to represent all clades in these two publications (where available, protein sequences were directly taken from GenBank annotations; otherwise, produced by Prodigal v2.6.3 [49], trained on the set of 12 complete or chromosome-level assembled Asgard genomes). To only include those sequences which are widely present within prokaryotic families, and to minimize the possibility of post LECA horizontal transfer from eukaryotes, soft-core pangenomes were constructed for each of our 26 curated taxonomic groups, mostly conforming to the NCBI taxonomy rank "class" (see below). To construct the soft-core pangenomes, all sequences from each taxonomic group were clustered individually using mmseqs2 "mmseqs cluster -s 7 -c 0.9 --cov-mode 0" [31]. All clusters which did not contain sequences from more than 67% of all bacterial families within the prokaryotic class were rejected from the database.

The eukaryotic database was constructed starting from a curated set of 72 genomes available on the NCBI GenBank. In order to provide a more accurate representation of eukaryotic diversity, this dataset was supplemented with EukProt v3 [33], a curated database aiming to provide a sparse representation of eukaryotic biology, to create Euk72Ep containing ~30.3 million sequences. As sequences included in EukProt v3 come from a diverse range of sources, some known to contain contaminant prokaryotic sequences, the database was screened for prokaryotic contamination. This screen was carried out by first constructing candidate prokaryotic and eukaryotic HMM databases as described below and querying both databases against the eukaryotic sequence database using hhblits [32]. All eukaryotic sequences with a top hit alignment score from a prokaryotic HMM were removed from the database prior to initial clustering.

Curation of taxonomic labels

In order to provide a compromise between taxonomic specificity and accuracy of clade assignment, we curated a set of taxonomic labels for both Euk72Ep and Prok2111As to act as representatives. To construct this set, we started with manually assigning all species in EukProt to their closest relative in the NCBI taxonomy. Then, each species from Prok2111As and Euk72Ep was assigned a taxonomic label corresponding to their "Class" rank (for example, Alphaproteobacteria, Thermococci, Mammalia) based on the NCBI Taxonomy of November 2023 using tools from the ete3 3.1.3. toolkit [50] as well as custom scripts. For species without a corresponding "Class" rank, the closest relevant rank was manually assigned. This provided an initial list of 317 unique taxonomic class identifiers.

516 As the rank "Class" does not evenly partition biological diversity, identifiers were further
517 manually mapped to a set of 45 eukaryotic and 26 prokaryotic taxonomic superclasses, each
518 present in the NCBI taxonomy (for example, Metazoa, Streptophyta, TACK archaea) and
519 described in (supplementary data Y). Due to the partially unresolved taxonomic relationship
520 with Asgard archaea, all candidate Asgard sequences from Prok2111 or from Refs. 19,20 were
521 assigned the class label "Asgard".
522
523 **Sequence clustering and profile database generation**
524 To transform Euk72Ep and Prok2111As into profile databases suitable for sensitive HMM-HMM
525 searches, we implemented an unsupervised, cascaded sequence-profile clustering pipeline using
526 the tools available in the mmseqs2 software suite [31]. In brief, each sequence database was
527 initially clustered at 90% sequence identity and 80% pairwise coverage using "mmseqs linclust"
528 and cluster representatives were chosen using "mmseqs result2repseq". This procedure
529 provides a non-redundant set of 18.6 million prokaryotic and 25.1 million eukaryotic sequences
530 for further analysis. The non-redundant sequences are further collapsed into a set of initial
531 profiles and consensus sequence pairs using "mmseqs cluster", "mmseqs result2profile" and
532 "mmseqs profile2consensus". All consensus sequences are queried against their profiles using
533 "mmseqs search" and results clustered using "mmseqs clust". Clusters are mapped to their
534 original non-redundant sequences using "mmseqs mergeclusters" and profiles and consensus
535 sequences are constructed once again. In order to grow clusters based on their sequence-
536 profile alignments, this procedure of "mmseqs profile2consensus - mmseqs search - mmseqs
537 clust - mmseqs result2profile" is iterated until cluster sizes converge. An 80% pairwise coverage
538 is maintained throughout all steps of the cascaded clustering protocol. The final clustered
539 databases contain 14.1 million and 1.7 million clusters respectively for Euk72Ep and Prok2111As
540 (supplementary Figure 1).
541
542 To transform the unsupervised clusters into a set of HMM profiles, we utilized a mixed MSA
543 construction strategy coupled with automatic data reduction to automatically create accurate
544 alignments for a wide diversity of cluster sizes and sequence diversities. To construct HMMs, all
545 non-singleton sequence clusters are first aligned using FAMSA 2.0.1. [51] to produce an initial
546 alignment. Using the initial alignment, we then reduce the sequence space by first creating an
547 approximate tree using FastTree 2.1.1. using "FastTree -gamma" [52] and iteratively removing the
548 closest pairwise leaves using ete3 [50], keeping the leaf closest to the root, until the desired
549 sequence set size is reached. For profile generation, we retain no more than 300 sequences for
550 both Prok2111As and Euk72Ep. This reduced sequence set is then realigned with muscle5
551 5.1.linux64 [36] using "muscle -diversified" with 5 replicates, and the maximum column confidence
552 alignment is extracted. This alignment is then trimmed to keep columns with more than 0.2 bits
553 of Shannon information. This process of data reduction/alignment/trimming is referred to as
554 "prune-and-align" and employed later for downstream analysis. Trimmed alignments are turned
555 into an HMM database using HHsuite with "hhconsensus -M 50", hhmake and "cstranslate -f -x
556 0.3 -c 4 -I a3m". The final HHsuite databases contained 480.000 and 1.630.000 profiles for
557 prokaryotes and eukaryotes, respectively, see (supplementary Figure 1).

558 **Search for prokaryotic homologs of eukaryotic proteins**
559 Given that clade-specific protein families were not of direct interest to this study, we reduced
560 the search space by excluding HMM profiles from eukaryotic clusters with less than 10
561 sequences and a lowest common ancestor with a taxonomic rank below "Superkingdom" as
562 calculated by "mmseqs lca". This requires a sequence cluster to contain at least two sequences
563 from different kingdoms, as defined by the NCBI taxonomy (ex. Amoebozoa and Haptista, or
564 Viridiplanta and Opistokonta), retaining 142.000 profiles of conserved eukaryotic proteins. To
565 associate eukaryotic sequence clusters with homologous prokaryotic clusters, we queried the
566 Euk72Ep HMM database against all Prok2111As HMM profiles using a single iteration of HHBlits
567 3.3.0 as "HHblits -n 1 -p 80" and retaining the hits with at least 80% HHblits probability and
568 80% pairwise profile length. These filters resulted in a total of 36,673 accepted eukaryotic
569 query profiles targeting a total of 36634 unique prokaryotic profiles. These clusters contain a
570 total of 10.4 million and 2.7 million sequences, mapping to 55 % and 10 % of original
571 prokaryotic and eukaryotic sequences, respectively.
572
573 **EPOC MSA generation**
574 Unless otherwise stated, all subsequent steps were carried out using custom python 3.9 scripts,
575 and tree structures and taxonomy were handled using ete3 [50]. All sequences from a eukaryotic
576 cluster with at least 10 sequences and a lowest common ancestor at the rank of
577 "Superkingdom" (NCBI taxonomy), together with all members of the homologous prokaryotic
578 clusters identified in the HMM-HMM search with at least 80% probability and pairwise sequence
579 coverage, forms a EPOC. As EPOCs varied in size by more than 5 orders of magnitude (10 to
580 100.000 sequences), robust data subsampling is necessary for accurate downstream MSA
581 generation and tree construction. We employed a variant of the "prune-and-align" strategy
582 described above for profile generation taking into account the taxonomic distribution of the
583 constituent sequences. Rather than iteratively pruning the closest leaves pairwise, we first label
584 each leaf with its corresponding taxonomic label and only prune monophyletic groups,
585 maintaining a count of all pruned leaves per clade. If the procedure fails to reach the target leaf
586 number before converging, all leaves are ordered by the number of pruned relatives and leaves
587 with the lowest number of relatives deleted (retaining the largest monophyletic groups) until
588 the target size is reached. The result approximates a maximum diversity representation,
589 preferentially pruning isolated singleton clades. Using this modified protocol, eukaryotic and
590 prokaryotic sequences are cropped to a maximum of 30 and 70 sequences, respectively. Each
591 EPOC therefore consists of a maximum of 100 representative sequences. All EPOCs are aligned
592 using "muscle -diversified" with 5 replicates, and the maximum confidence alignment is
593 extracted using the --maxcc argument. This alignment is then trimmed to keep columns with
594 more than 0.15 bits of Shannon information content for tree reconstruction.
595
596
597
598
599

EPOC tree construction and processing

From the pruned and trimmed alignment, we created a maximum likelihood tree using IQtree2 2.3.5. "IQtree2 -B 1000 -bnni -m MFP -mset LG,Q.pfam --cmin 4 --cmax 12" with model parameters estimated by model finder plus [37]. Candidate models and rate category search ranges were selected based on a test set of 200 randomly chosen EPOCs from the filtered core set analyzed with full model parameter evaluation by Model finder [37] and 10 tree replicates. The LG and Q.pfam models were chosen as they consistently represented the highest ranking log-likelihoods. Rate categories were assigned based on the highest and lowest amount found among the top 10 average scoring models from the same set.

As the master trees are built from sequences obtained through sensitive cascaded sequence-profile clustering we would in rare cases observe long-stemmed clade outliers, in addition to more common long stem leaf outliers. In order to assess and remove such erroneous data, we first estimated the log-normal distribution of all stem lengths using scipy.stats 1.11.1 and excluded any stems outside the 0-99.5 % Probability Point Function interval. This simultaneously removes long terminal leaf branches as well as highly diverged clades. In rare cases where this pruning would remove either the entire eukaryotic clade or more than 30 % of all leaves, the EPOC was discarded (387 EPOCs). The resulting trees were then re-rooted using a weighted midpoint approach so that the sum of all branch stem lengths on either side of the root was equal.

To assign taxonomic clades, all leaf nodes were assigned taxonomic labels from our curated set and monophyletic clades were collapsed, taking as the representative the lowest branching leaf. We found that assigning clades purely based on a monophyletic definition of taxonomic purity severely hampered the analysis as leaves can be taxonomically incongruous with their neighbors. This effect most likely stems from cases of locally erroneous tree topology or local HGT, but severely complicates the downstream analysis. We therefore adopted the notion of a "soft LCA" representing the root of a close-to-monophyletic clade. To greedily rank all the best "soft LCAs", for each taxonomic label, the tree was traversed from each monophyletic clade root to the tree root, calculating a score for each internal node;

(number of leaves with label X in clade / clade size) * (number of leaves with label X in clade / total number of label X)

This metric balances taxonomic "purity" and "scope" for each possible clade of label X. All such nodes are then ordered by score. This allows for paraphyly across all taxonomic labels. To avoid overinterpreting small clades, we only consider clades to be valid if they represent at least 3 sequences with clade purity > 0.8 for prokaryotes, and at least 5 sequences with purity > 0.8 for eukaryotes. Trees that fail to identify either any eukaryotic or any prokaryotic clades under these constraints, primarily derived from small EPOCs, are discarded (12113 EPOCs).

642 **Evolutionary hypothesis testing using constraint trees**
643 In order to assess the relative probabilities of each sampled prokaryotic "soft LCA" to represent
644 the eukaryotic sister clade, we employed evolutionary hypothesis testing using constraint trees
645 as implemented in IQtree2 [37]. For each EPOC and for each detected eukaryotic clade, as
646 defined by its corresponding "soft LCA", we generated a set of constraint trees, one for each
647 possible prokaryotic sister. Each constraint tree had three defined clades, 1) the eukaryotic
648 group as defined by all sequences below the eukaryotic LCA, 2) the prokaryotic sister group and
649 3) all other prokaryotic leaves. We then construct a set of local trees from slices of the original
650 MSA using IQtree2, forcing the constraint tree topology using "iqtree2 -g", using the same
651 evolutionary model as the original, unconstrained "master tree". The set of constrained trees is
652 then ranked using "iqtree2 -z" calculating the relative model assignment confidence for each of
653 the constrained topologies and resulting test metrics are saved for downstream analysis. As the
654 number of trees to evaluate scales by the number of eukaryotic clades multiplied by the number
655 of prokaryotic clades, we limit the sampling to a maximum of 3 clades per tree for eukaryotes,
656 discarding rare cases of high eukaryotic paraphyly, and only consider the 12 closest prokaryotic
657 clades to each individual eukaryotic clade, based on their topological distance (supplementary
658 Figure 2). The procedure results in an Expected Likelihood Weight for each of the evaluated
659 model trees, here taken as analogous to model selection confidence that an evaluated
660 prokaryotic clade is the most likely sister to a eukaryotic clade.
661
662 **EPOC annotation**
663 To functionally annotate protein families within EPOCs, we generated profiles based on protein
664 sequences in KEGG release 110 [35]. For each Kegg Orthologous Group (KOG), we extracted Kegg
665 metadata including KEGG pathways and BRITE labels via KEGG's API. At the time of parsing,
666 KEGG contained 26,695 KOGs. Due to KEGG API constraints, protein sequences in KEGG were
667 extracted from Uniprot, first by mapping KEGG proteins to Uniprot IDs via Unirprot's ID
668 mapping service (uniprot.org/id-mapping), then by extracting the sequences for these Uniprot
669 IDs corresponding to each KOG [53].
670
671 For each KOG, we generated HHM profiles using the aforementioned prune-and-align pipeline.
672 To increase specificity of labeling, we partitioned those KOGs which encompass both prokaryotic
673 and eukaryotic sequences into separate taxonomic groups and sequence alignments were
674 generated separately for prokaryotes and eukaryotes. Each eukaryotic HMM profile forming the
675 basis for a EPOC was queried against our KEGG profile database using "HHblits -n 3 -p 80".
676 Results were subsequently filtered to a pairwise profile coverage of 0.5 using custom scripts. As
677 a result, 17,176 of the 18,834 EPOCs could be annotated by at least one KOG profile at 80%
678 probability and 0.5 pairwise coverage. For all downstream analyses, we considered only the top
679 hit as relevant for annotation.
680
681
682
683

Data filtering and removal of likely horizontal gene transfer cases

Unless otherwise stated, all data is interpreted from the core set of EPOCs meeting the following criteria: 1) the eukaryotic constituent cluster profile identifies at least one target profile in our KEGG profile database at a probability of 80% and 50% coverage, 2) the eukaryotic clades include more than 5 distinct taxonomic labels, at least one from both Amorphea and Diaphoretickes, thus encompassing LECA, 3) only prokaryotic sister clades with $0.4 < ELW < 0.99$ were considered reliable and were included in the analysis. Cases of $ELW < 0.99$ were not considered as these are found to be disproportionately eukaryotic-like, often branching from within the eukaryotic clade itself, and thus likely derived from late horizontal transfer post LECA. Due to the presence of such high ELW values within Alphaproteobacterial associations with Oxidative phosphorylation, in this case, as an exception, ELW values of 1 were included.

Data visualization and plotting

All plots were prepared using python v3.9 and pandas v.2.0.3 with Altair v.4.2.2. Layout, annotation, and vector editing was done using Inkscape v.1.1.1. All statistical tests were carried out using scipy.stats 1.11.1

References

1 Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623-630 (2006).

2 Vosseberg, J. *et al.* The emerging view on the origin and early evolution of eukaryotic cells. *Nature* **633**, 295-305 (2024). https://doi.org:10.1038/s41586-024-07677-6

3 Lopez-Garcia, P. & Moreira, D. The symbiotic origin of the eukaryotic cell. *C R Biol* **346**, 55-73 (2023). https://doi.org:10.5802/crbiol.118

4 Gabaldon, T. Origin and Early Evolution of the Eukaryotic Cell. *Annu Rev Microbiol* **75**, 631-647 (2021). https://doi.org:10.1146/annurev-micro-090817-062213

5 de Duve, C. The origin of eukaryotes: a reappraisal. *Nat Rev Genet* **8**, 395-403 (2007). https://doi.org:10.1038/nrg2071

6 Alberts, B. *et al. Molecular Biology of the Cell*. 7 edn, (W.W. Norton & Co, 2022).

7 Cavalier-Smith, T. Eukaryotes with no mitochondria. *Nature* **326**, 332-333 (1987). https://doi.org:10.1038/326332a0

8 Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & Peattie, D. A. Phylogenetic meaning of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia. *Science* **243**, 75-77 (1989). https://doi.org:10.1126/science.2911720

9 Sagan, L. On the origin of mitosing cells. *J Theor Biol* **14**, 255-274 (1967).

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

11 Lopez-Garcia, P. & Moreira, D. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem Sci* **24**, 88-93 (1999). https://doi.org:10.1016/s0968-0004(98)01342-5

12 Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61**, 456-502 (1997). https://doi.org:10.1128/mmbr.61.4.456-502.1997

13 Koonin, E. V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* **5**, R7 (2004).

14 Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* **95**, 6239-6244 (1998). https://doi.org:10.1073/pnas.95.11.6239

15 Ribeiro, S. & Golding, G. B. The mosaic nature of the eukaryotic nucleus. *Mol Biol Evol* **15**, 779-788 (1998). https://doi.org:10.1093/oxfordjournals.molbev.a025983

16 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). https://doi.org:10.1099/ijsem.0.006435

17 Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173-179 (2015). https://doi.org:10.1038/nature14447

18 Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of eukaryotes. *Nat Rev Microbiol* **15**, 711-723 (2017). https://doi.org:10.1038/nrmicro.2017.133

19 Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353-358 (2017). https://doi.org:10.1038/nature21031

20 Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553-557 (2021). https://doi.org:10.1038/s41586-021-03494-3
21 Eme, L. *et al.* Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992-999 (2023). https://doi.org:10.1038/s41586-023-06186-2
22 Makarova, K. S. *et al.* Diversity, origin, and evolution of the ESCRT systems. *mBio* **15**, e0033524 (2024). https://doi.org:10.1128/mbio.00335-24
23 Lombard, J., Lopez-Garcia, P. & Moreira, D. The early evolution of lipid membranes and the three domains of life. *Nat Rev Microbiol* **10**, 507-515 (2012). https://doi.org:10.1038/nrmicro2815
24 Moreira, D. & Lopez-Garcia, P. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J Mol Evol* **47**, 517-530 (1998). https://doi.org:10.1007/pl00006408
25 Lopez-Garcia, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655-667 (2020). https://doi.org:10.1038/s41564-020-0710-4
26 Krupovic, M., Dolja, V. V. & Koonin, E. V. The virome of the last eukaryotic common ancestor and eukaryogenesis. *Nat Microbiol* **8**, 1008-1017 (2023). https://doi.org:10.1038/s41564-023-01378-y
27 Lopez-Garcia, P. & Moreira, D. Cultured Asgard Archaea Shed Light on Eukaryogenesis. *Cell* **181**, 232-235 (2020). https://doi.org:10.1016/j.cell.2020.03.058
28 Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519-525 (2020). https://doi.org:10.1038/s41586-019-1916-6
29 Gabaldon, T. Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial symbioses" hypothesis. *IUBMB Life* **70**, 1188-1196 (2018). https://doi.org:10.1002/iub.1950
30 Pittis, A. A. & Gabaldon, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic ancestry. *Nature* **531**, 101-104 (2016). https://doi.org:10.1038/nature16941
31 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017). https://doi.org:10.1038/nbt.3988
32 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019). https://doi.org:10.1186/s12859-019-3019-7
33 Richter, D. J. *et al.* EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes *Peer Community J* **2**, e56 (2022).
34 Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc Biol Sci* **269**, 137-142 (2002). https://doi.org:10.1098/rspb.2001.1862
35 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017). https://doi.org:10.1093/nar/gkw1092
36 Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022). https://doi.org:10.1038/s41467-022-34630-w
37 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). https://doi.org:10.1093/molbev/msaa015
38 Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci U S A* **112**, E693-699 (2015). https://doi.org:10.1073/pnas.1420657112

39 Al Jewari, C. & Baldauf, S. L. Conflict over the Eukaryote Root Resides in Strong Outliers, Mosaics and Missing Data Sensitivity of Site-Specific (CAT) Mixture Models. *Syst Biol* **72**, 1-16 (2023). https://doi.org:10.1093/sysbio/syac029

40 Richards, T. A. & van der Giezen, M. Evolution of the Isd11-IscS complex reveals a single alpha-proteobacterial endosymbiosis for all eukaryotes. *Mol Biol Evol* **23**, 1341-1344 (2006). https://doi.org:10.1093/molbev/msl001

41 Kassube, S. A. & Thoma, N. H. Structural insights into Fe-S protein biogenesis by the CIA targeting complex. *Nat Struct Mol Biol* **27**, 735-742 (2020). https://doi.org:10.1038/s41594-020-0454-0

42 Hedges, S. B. The origin and evolution of model organisms. *Nat Rev Genet* **3**, 838-849 (2002). https://doi.org:10.1038/nrg929

43 Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol* **5**, 92-100 (2021). https://doi.org:10.1038/s41559-020-01320-z

44 Caforio, A. *et al.* Converting Escherichia coli into an archaebacterium with a hybrid heterochiral membrane. *Proc Natl Acad Sci U S A* **115**, 3704-3709 (2018). https://doi.org:10.1073/pnas.1721604115

45 Hoekzema, M., Jiang, J. & Driessen, A. J. M. Optimizing Archaeal Lipid Biosynthesis in Escherichia coli. *ACS Synth Biol* **13**, 2470-2479 (2024). https://doi.org:10.1021/acssynbio.4c00235

46 Tamarit, D. *et al.* A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal viruses. *Nat Microbiol* **7**, 948-952 (2022). https://doi.org:10.1038/s41564-022-01122-y

47 Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* **613**, 332-339 (2023). https://doi.org:10.1038/s41586-022-05550-y

48 Munoz-Gomez, 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). https://doi.org:10.1038/s41559-021-01638-2

49 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010). https://doi.org:10.1186/1471-2105-11-119

50 Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**, 1635-1638 (2016). https://doi.org:10.1093/molbev/msw046

51 Deorowicz, S., Debudaj-Grabysz, A. & Gudys, A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Sci Rep* **6**, 33964 (2016). https://doi.org:10.1038/srep33964

52 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010). https://doi.org:10.1371/journal.pone.0009490

53 UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-D489 (2021). https://doi.org:10.1093/nar/gkaa1100