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

The origin of eukaryotes is one of the key problems in evolutionary biology [1,2]. 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 [3-7], inform and constrain evolutionary scenarios of eukaryogenesis [8]. We undertook a comprehensive analysis of the origins of 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 clear consistent trends. These findings suggest a model of eukaryogenesis in which key features of eukaryotic cell organization evolved in the Asgard lineage leading to the LECA, 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.

Main

Eukaryotes drastically differ from archaea and bacteria (collectively, prokaryotes) by the highly complex organization of the eukaryotic cells. The signature features of this organizational complexity include the eponymous nucleus, the endomembrane system, the elaborate cytoskeleton, and the energy-converting mitochondrion which evolved from an Alphaproteobacterial endosymbiont [9]. Early models of eukaryogenesis featured a protoeukaryote lacking mitochondria that captured and domesticated an alphaproteobacterium at a relatively late stage of evolution [10,11]. However, subsequent research revealed no primary amitochondrial eukaryotes (although many have lost mitochondria through reductive evolution) [12], suggesting that the Last Eukaryotic Common Ancestor (LECA) was already endowed with mitochondria along with the other signatures of the eukaryotic cellular organization [1,2,8]. These findings lend credence to scenarios of eukaryogenesis in which mitochondrial endosymbiosis triggered cellular reorganization, giving rise to the eukaryotic cellular complexity [1,2,8,13-15].

Phylogenomic analyses show that those conserved, core genes of eukaryotes that can be traced to the LECA are a mix originating from archaea and various bacteria, and in early studies, the bacterial contributions quantitatively exceeded the archaeal ones. The archaea-derived genes were found primarily in information processing systems (replication, transcription and translation) whereas genes of apparent bacterial origin comprised the operational component of the eukaryotic gene complement, in particular encoding metabolic enzymes [16-19].

The study of eukaryogenesis was transformed by the discovery and exploration of Asgard archaea (currently, phylum *Promethearchaeota* within kingdom *Promethearchaeati* [20, 21], hereafter Asgard) which includes the closest known archaeal relatives of eukaryotes [3-7]. Phylogenetic analyses of highly conserved genes have been pushing the eukaryotic branch progressively deeper within the Asgard tree. The latest such analysis identified the order *Hodarchaeales*, within the Asgard class *Heimdallarchaeia*, as the likely sister group of eukaryotes [4]. However, with the continued influx of Asgard sequence data, the exact affinity of eukaryotes within Asgard remains a moving target [22,23]. Notably, it has been shown that Asgard archaea encode, express and utilize homologs of a broad variety of eukaryote signature proteins beyond the core information processing componentry; in particular, cytoskeletal proteins and proteins involved in membrane remodeling [4,5,7,24,25]. Furthermore, these genes of Asgard archaea have undergone extensive duplication and subfunctionalization pre-LECA, further highlighting their prominent role in eukaryogenesis [26].

A variety of models for eukaryogenesis have been proposed, differing with respect to the timing, topology and origin of the eukaryotic cellular organization and genetic repertoire [1,2,8]. 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 host [2,8,27]. However, a major conundrum for such scenarios of eukaryogenesis is the chemistry of membrane lipids and the enzymology of their biosynthesis, which are unrelated in archaea and bacteria, with eukaryotes utilizing the bacterial type [28]. Accordingly, any model of eukaryogenesis that includes an archaeal host would require a change in membrane composition. More complex, alternative models of eukaryogenesis, underpinned by metabolic symbiosis, postulate two endosymbiotic events whereby an Asgard archaeon was first engulfed by a bacterium, the archaeal membrane being lost in the process, and then, a second endosymbiosis gave rise to the mitochondria [15,29,30]. Even if seemingly less parsimonious than binary scenarios, such models of eukaryogenesis account for the continuity of bacterial membranes (while requiring retargeting at least part of the Archaeal membrane proteome) and appear compatible with the syntrophic lifestyle of at least some Asgard archaea, found in consortia with bacteria, Myxococcota in particular (formerly Deltaproteobacteria)[31,32]. Furthermore, a pivotal study by Pittis and Gabaldon [35] 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 mitochondrion, apparently preceded by acquisition of many genes from different bacteria, possibly through one or multiple earlier endosymbioses [33-35].

Here, we took advantage of the rapidly growing collection of archaeal, bacterial and eukaryotic genome sequences to assess the origins of core eukaryotic genes within a rigorous statistical framework based on evolutionary hypotheses testing using constrained phylogenetic trees. The results reveal a consistent, major link between Asgard archaea and the origin of most functional classes of eukaryotic genes, demonstrating their dominant contribution to eukaryogenesis. These findings are compatible with a scenario of eukaryogenesis where many signature, complex features of eukaryotic cells evolved in the Asgard ancestor of eukaryotes.

Phylogenomic analysis and global evolutionary associations of conserved eukaryotic proteins

To represent prokaryotes, a database of 75 million prokaryote protein sequences was curated from 47,545 complete prokaryotic genomes obtained from the NCBI GenBank in November 2023 (prok2311) and supplemented with proteins extracted from 63 Asgard genome assemblies [4,5]. The initial eukaryotic database consisted of 30 million sequences from 993 species present in EukProt v3 [73], cleaned using mmseqs2 [75] to remove prokaryotic contaminants (Figure 1A). In order to accurately infer the origins of genes in the LECA, it is essential to exclude genes present only within narrow subsets of species. These possibly result from post-LECA horizontal gene transfer (HGT) between eukaryotes and prokaryotes, a major phenomenon given the pervasive cohabitation of eukaryotes with prokaryotes [36-39]. To prevent such genes from affecting our analysis, we reconstructed the "soft-core" pangenome for each of our assigned prokaryotic and eukaryotic taxonomic classes (see Methods). These pangenomes included only those genes that are present in at least 50% of the species within each class of bacteria and archaea in prok2311 (20.7 million), and at least 20% of the species in eukaryotes (12.7 million), ensuring that our inferences were based only on widespread protein families and would not reflect lineage-specific HGT (supplementary Figure 1D).

To identify close connections 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 based on mmseqs2 clustering, followed by a multistep data-reduction and multiple sequence alignment (MSA) procedure to generate HMM profiles using diversified muscle5 alignments [43] and HH-suite [40] (see Methods). The resulting eukaryotic HMMs were queried against the prokaryotic dataset using HHblits [40] to identify sets of homologous protein sequences. Each eukaryotic cluster sequence and all its significant prokaryotic hits constituted an individual set, hereinafter referred to as a Eukaryotic-Prokaryotic Orthologous Cluster (EPOC). Each EPOC contains a unique set of eukaryotic proteins. These EPOCs were used for phylogenetic tree construction, annotation, and evolutionary hypothesis testing (Figure 1A). The final EPOCs include 10.9 million prokaryotic and 1.74 million eukaryotic sequences, mapping to 52% and 14% of the respective non-redundant pangenome datasets.

To infer the most likely prokaryotic ancestry of the eukaryotic proteins given the data in each EPOC, rather than relying on tree topology directly, we employed a probabilistic approach for evolutionary hypothesis testing using constraint trees (Figure 1A). Following the construction of an initial master tree, we carried out further constrained tree calculations, exhaustively sampling all arrangements of the prokaryotic sister clades nearest to the eukaryotic clade(s) and obtaining Expected Likelihood Weights (ELW) for the set of possible sister clade models (supplementary Figure 2, supplementary Data 1) [41]. Given that the ELW metric is analogous to model selection confidence, here we take it to be proportional to the probability that a sampled prokaryotic clade is the best estimate of the closest sister clade of a eukaryotic clade. For each EPOC, our analysis dynamically accounts for long branch outliers and is capable of resolving eukaryotic paraphyly, treating each resolved eukaryotic clade within a EPOC individually for downstream analysis (see Methods). The resulting data included 14,400 EPOCs

131 annotated using profiles generated from KEGG Orthology Groups (KOGs) [42], each with an MSA
132 generated using muscle5 [43], a maximum likelihood tree inferred using IQtree2 [44] and associated ELW
133 values for all candidate prokaryotic sister clades. The analysis of prokaryotic ancestry was performed
134 only for those eukaryotic clades that included more than 5 distinct taxonomic labels, with at least one
135 coming from Amorphea and one from Diaphoretickes, the two expansive eukaryotic clades considered
136 to emit from either the first or the second bifurcation in the evolution of eukaryotes [45,46]. Thus, although
137 we did not attempt to explicitly reconstruct the LECA gene set, these clades represent genes which,
138 given their broad taxonomic distribution in our data, were likely to have had an ancestor in the LECA.
139
140 Considering the global distribution of ELW values across all 13,500 EPOCs covering 3950 unique KOGs,
141 the single greatest average ELW (aELW), here referred to as "association", was with Asgard archaea.
142 Further associations with Cyanobacteria, Actinomycetota, Betaproteobacteria and Alphaproteobacteria,
143 as well as trace association with many additional bacterial classes, were detected at lower levels (Figure
144 1B). Due to the scope of the analysis, MSA quality and topological diversity of the trees varied across the
145 EPOCs such that some EPOCs showed low ELW values for all clades, indicating that the data in these
146 cases was insufficient to discern their eukaryotic ancestries. Excluding such low-quality EPOCs with a
147 maximum ELW < 0.4 yielded a robust core set of 4290 EPOCs covering 2100 KOGs, improving the
148 reliability of the results. This core set covers a wide range of information processing systems, ubiquitous
149 metabolic pathways, and transporters, as well as regulatory and housekeeping proteins. Limiting the
150 analysis to this core subset of well assigned eukaryotic families with wide taxonomic coverage increased
151 the global Asgard association, which now accounted for 50% of all likelihood weight across more than
152 4,000 unique data points, covering more than 2,000 unique KOGs (Figure 1B).
153
154 To ensure these results were robust to the choice of the clustering approach, we performed additional
155 cluster annealing steps (see Methods, supplementary Figure 8A). In brief, starting with the mmseqs2
156 clusters, we performed HMM-HMM searches and clustered the results with a greedy set cover
157 approach. We then further partitioned these superclusters by constructing per-cluster trees and
158 identifying sets of leaves with mutually short pairwise distances, indicating the existence of separate
159 clades. These annealed clades, now strictly defined to be monophyletic, given significant sequence
160 similarity, were then used for EPOC construction. Additionally, we reconstructed our full dataset using
161 the Genome Taxonomy Database (GTDB) taxonomy to verify that the detected eukaryote-prokaryote
162 associations were not substantially affected by our use of the NCBI taxonomy [47]. We note that, while
163 relative contributions from specific taxa varied due to remapping of taxonomic labels (for example,
164 merging of Betaproteobacteria into Gammaproteobacteria), or due to the stricter annealed EPOC
165 definition limiting the scope of the data; our key observation of a dominant Asgard association for
166 eukaryotes held independent of clustering and taxonomy (supplementary Figure 1E, 8B, 8C). Given the
167 robustness of the results to both cluster specification and taxonomy, all subsequent analysis was based
168 on the core EPOC set described above. Finally, although our results suggest the presence of diverse
169 paralogs of apparent Asgard origin in the LECA [5,26,48], paralogy cannot account for this wide contribution
170 of Asgard to eukaryogenesis (supplementary Figure 1B, supplementary Figure 3A).
171

Averaging ELW scores across EPOCs sharing KEGG ontology supported strong Asgard association with, among other functional systems and pathways, the ribosome, RNA and DNA polymerases, Ras-like GTPases, as well as ubiquitin-mediated protein degradation and the proteasome. These findings validated and expanded upon previous observations [2,4,5], extending Asgard associations to large parts of the core metabolic network (Figure 1C). In contrast, prominent Alphaproteobacterial associations were relatively limited, primarily including proteins involved in oxidative phosphorylation and Fe-S clusters biogenesis, with weaker associations across the mitochondrial metabolic pathways. Thus, we observed a far stronger overall association between Asgard archaea and eukaryotes across diverse biological functions and pathways than previously described [4,5] although the association between eukaryotic core metabolism and a wide range of bacterial phyla remained cumulatively substantial. We also observed a consistent association with Cyanobacteria, a known endosymbiont of Archaeplastida and the ancestor of plastids, with a strong impact on the evolution of Eukarya post LECA. However, we note that pathways driving this Cyanobacterial association are unrelated to the photosynthetic capabilities of the chloroplast, but rather exhibit a broad association with multiple metabolic pathways, without apparent trends, similar to the observations for other bacterial taxa (supplementary Figure 4)[49-51]. Thus, we conclude that this Cyanobacterial signal most likely stems from events pre-LECA and is distinct from the later contributions from the chloroplast.

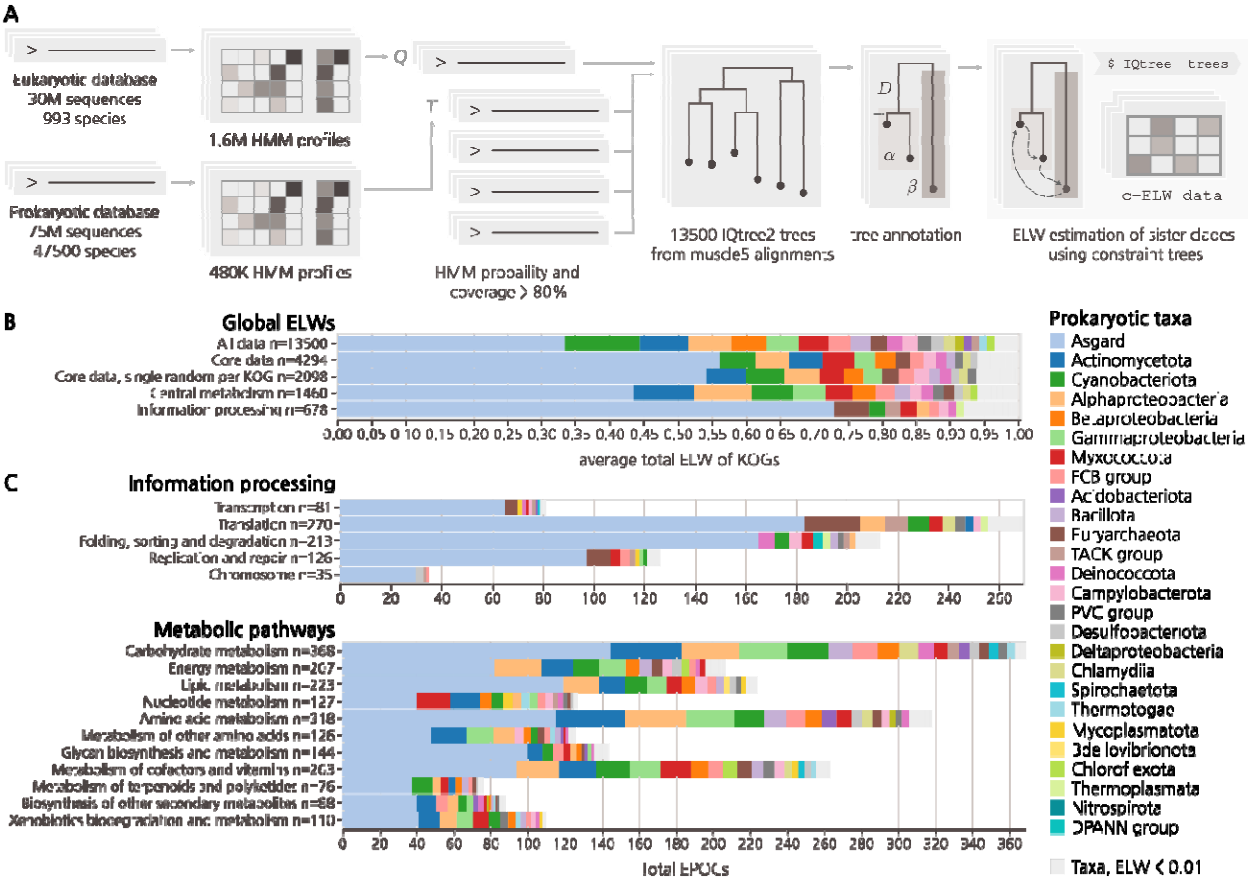


Figure 1. Phylogenetic associations of functional classes of eukaryotic proteins with major archaeal and bacterial taxa.

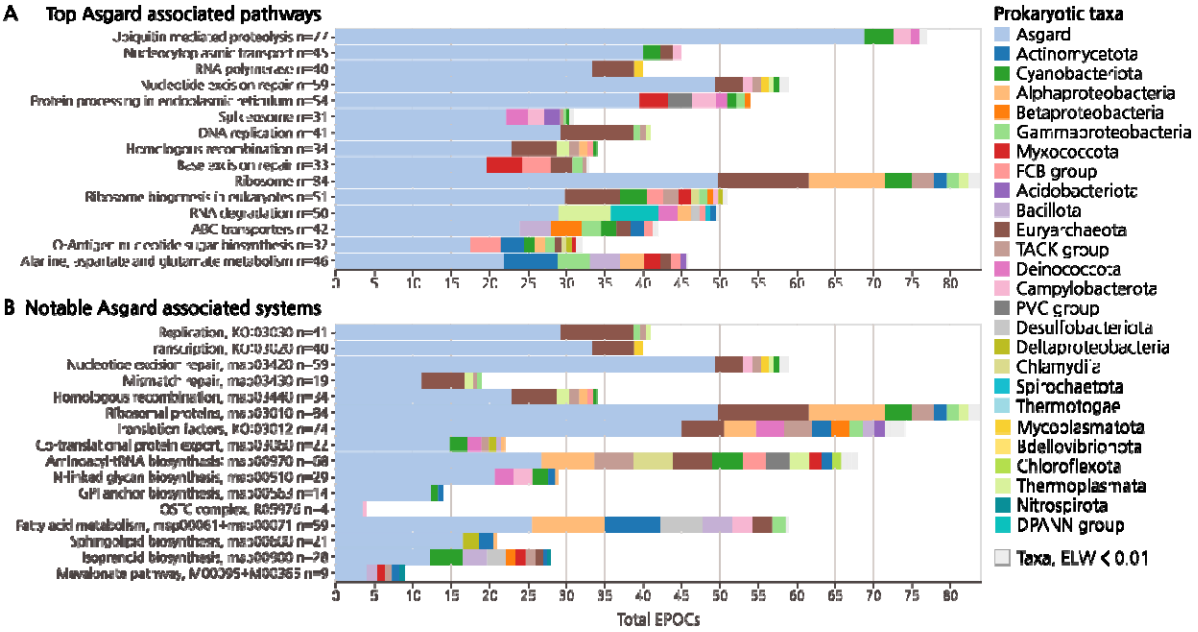
193 **A)** Simplified EPOC generation and processing pipeline. From left to right, one eukaryotic and one
194 prokaryotic sequence database transformed into HMMs. The query of eukaryotic against prokaryotic
195 HMMs generated 13,500 annotated EPOCs. The detected prokaryotic clades were assessed for their
196 probability to be the closest eukaryotic sister phyla by permutation testing using IQTree2 (see
197 supplementary Figure 2). **B)** Global aELW for all EPOCs within unfiltered, core and non-paralogous
198 subsets and top-level breakdowns for metabolic enzymes and proteins of core information processing
199 pathways as defined by the KEGG ontology BRITE at the coarsest A level (KO:09100, KO:09120). Grey
200 markers group prokaryotic taxa with an aELW < 0.01. **C)** Further subdivision of aELW values across more
201 specific functional categories of genes BRITE level B, counting the total number of EPOCs and colored by
202 relative aELW of taxa per category.
203

204 ## Broad Asgard contributions to eukaryogenesis

205 In accord with a dominant contribution of Asgard archaea to eukaryogenesis, we observed consistent
206 and major associations of Asgard proteins with a wide array of cellular functions. In previous work, the
207 strongest Asgard traces have been noted across the information processing systems, with unambiguous
208 protein associations with DNA and RNA processing as well as protein expression, trafficking and
209 signaling [2,4,5]. Here, we confirmed and extended Asgard associations with genome replication and
210 transcription, and further detected pronounced Asgard traces for nucleotide excision repair, mismatch
211 repair and homologous recombination. Additional well-known associations such as ribosomal proteins
212 were expanded upon to include translation factors, components of the co-translational insertion
213 machinery, protein targeting and aminoacyl-tRNA biosynthesis (Figure 2A, supplementary data 1). Thus,
214 all groups of core eukaryotic proteins involved in information processing appear to be almost exclusively
215 of Asgard descent (Figure 1C).
216

217 We further detected Asgard associations extending far beyond the information processing systems,
218 including prominent contributions of Asgard archaea to the protein machinery involved in
219 nucleocytoplasmic transport as well as downstream protein sorting, glycosylation and targeting. Central
220 components of the ER-associated, N-linked glycan biosynthesis and transfer, including both cytoplasmic
221 and lumenal monoglycosyltransferases, as well as the core of the oligosaccharyltransferase complex
222 (OSTC) were strongly associated with Asgard (Figure 2A). Enzymes associated with glycosylation
223 maturation in the Golgi complex were not recovered in our core data, possibly, due to extensive
224 diversification of domain architecture in eukaryotes, limiting their taxonomic scope and apparent
225 presence in the LECA. The Asgard connections of the eukaryotic glycosylation machinery further
226 included the synthesis of GPI-anchors, which post-translationally tether targeted proteins to the
227 membrane [52], here detected as unambiguously Asgard-derived. We also detected an Asgard origin of the
228 7-subunit (UDP-GlcNAc)-transferring (GPI-GnT)-monoglycosyltransferase complex responsible for
229 initiating GPI-anchor synthesis, components required for the maturation of the GPI-anchor, as well as
230 the transamidase complex and factors responsible for transfer of the protein onto the mature GPI-
231 anchor (Figure 2B).
232

233 Of major importance to eukaryogenesis is the provenance of the pathways for the biosynthesis of
234 bacterial-type lipids, given that all binary symbiogenesis scenarios with a Archaeal host require a
235 transition from archaeal to bacterial lipids in the membranes [8]. Although Asgard associations were
236 observed for large parts of the overall metabolic network, the pathways associated with fatty acid
237 synthesis and decay showed a wide distribution of aELW values, suggesting a higher than average
238 degree of mosaicism. As we enforce a minimum ELW of 0.4 for the most likely ancestors in our core
239 data, such mosaicism is interpreted as separate, individually likely ancestries, averaged across EPOCs
240 within a pathway, rather than a general uncertainty of the constituent EPOCs. The global aELW values
241 favored Asgard origin for these pathways, but there were also associations with Actinomycetota and
242 Alphaproteobacteria, further discussed below. However, higher Asgard aELW was observed within the
243 adjoining ER localized pathways for sphingolipid metabolism, a broad class of derived plasma membrane
244 lipids in eukaryotes. Previous studies have highlighted possible convergent origins of this pathway in
245 bacteria and eukaryotes [53,54], but here we detected a global association with Asgard archaea (Figure 2A).
246 Of further note are the ER-associated mevalonate pathway, converting Acetyl-CoA to mevalonate and
247 further to Farnesyl and Geranyl diphosphate, which we also found to be associated with Asgard (Figure
248 2B). In eukaryotes, mevalonate is required for isoprenoid synthesis, in turn forming the precursor units
249 for both sterols, carotenoids and terpenoids, synthesized in the ER lumen via the mevalonate pathway.
250 However, in Archaea, isoprenoids are the precursors for the ether-linked membrane lipids [55]. In
251 conclusion, we detected Asgard associations across a wide range of cellular functions and metabolic
252 pathways and noted a relationship between Archaeal lipid synthesis and eukaryotic steroid precursor
253 metabolism. A distinctly weaker Asgard signal was detected for pathways involved in bacterial lipid
254 biosynthesis, suggesting a complex evolutionary history.
255



256
257 **Figure 2. Asgard origins of eukaryotic pathways. A)** Top 15 KEGG pathways with strongest Asgard aELW.
258 KEGG maps with more than 20 EPOCs included. Taxonomy colored as per Figure 1. **B)** Further notable

Asgard systems, aELW calculated over either KEGG "map" number (ex. map0000X), reaction number (R0000X), or module number (ex. M0000X).

Limited and specific Alphaproteobacterial contributions

In line with the central role of mitochondria in eukaryotic energy metabolism, and as recently reported for central carbon metabolism [56], we primarily observe clear associations between Alphaproteobacteria and mitochondrially localized metabolic pathways. 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 3B). 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 reported [57], the ISC assembly machinery is of Alphaproteobacterial origin, and accordingly, we found the 4Fe-4S ICSA platforms, as well as IBA57 and Fe-S cluster binding ferredoxin-1 and 2, to be associated with Alphaproteobacteria (Figure 3, supplementary Figure 3B, supplementary Data 1). However, the 2Fe-2S precursor scaffold ISCU showed a mosaic association with diverse prokaryotes (including Alphaproteobacteria), and a weaker but clearly detectable Asgard association. Furthermore, the cysteine desulfurase NFS1 and the upstream pathways for the biosynthesis of sulfur-containing amino acids, cysteine, and methionine, were Asgard associated. The Alphaproteobacterial associations with Fe-S synthesis were therefore restricted primarily to synthesis and integration of 4Fe-4S into the oxidative phosphorylation complexes. Outside the mitochondria, Alphaproteobacterial associations with Fe-S metabolism cease. Here, ISC insertion depends on the cytosolic targeting complex CIA, which consists of CIAO1, CIA2B and MMS19 [59]. For the CIA components CIA1 and CIA2B, we observed clear association with Asgard whereas MMS19 is not detected in our data. Taken together, these observations indicate that the contributions of Alphaproteobacteria to eukaryogenesis are mostly centered around the mitochondrion.

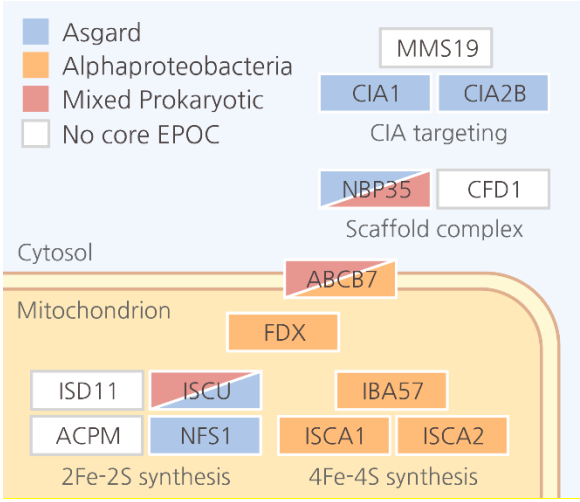


Figure 3. Ancestry of Eukaryotic Fe-S cluster synthesis pathways. Schematics for Fe-S cluster synthesis and assembly including downstream cytosolic Fe-S insertion are shown. Boxes indicate individual KOGs covered by EPOCs. Full color indicates support for ancestry with an aELW > 0.5, boxes with mixed color

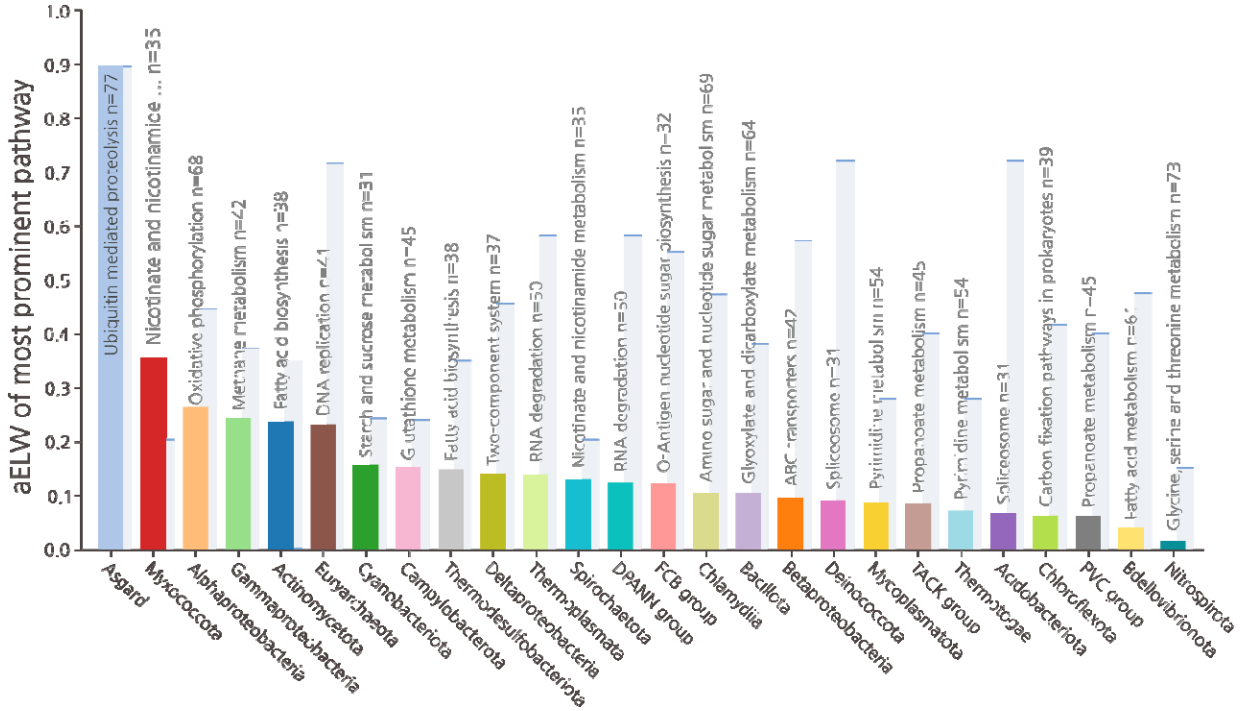
289 indicate unresolved ancestries with ELWs < 0.5 for the two most likely estimates. Mixed Prokaryotic
290 ancestries represent all possible taxa, including Alphaproteobacteria.
291

292 **Absence of functionally consistent contributions from other bacteria**

293 Although our analysis greatly expands the Asgard contributions to eukaryogenesis, and reveals limited
294 but consistent Alphaproteobacterial associations, contributions from other bacteria were detected as
295 well. For some biological functions, this "diverse" bacterial component accounted for the majority of the
296 aELW, and roughly one third of the analyzed KOGs (600 of 2,100) and EPOCs (1,580 of 4,300) had their
297 strongest association neither with Asgard, nor with Alphaproteobacteria . However, in a sharp contrast
298 to the limited but functionally related associations between Alphaproteobacterial and oxidative
299 phosphorylation and Fe-S metabolism, EPOCs associated with diverse bacteria showed few functional
300 trends, instead demonstrating low level association across most metabolic functions. Although
301 Cyanobacteria and Actinomycetota exhibited stronger global aELW compared to Alphaproteobacteria,
302 this association stemmed from multiple small contributions across disjoint pathways rather than specific
303 functionally related contributions (supplementary Figure 4). To emphasize this trend, across the diverse
304 set of bacterial taxa, only Alphaproteobacteria and Myxococcota were associated with a KEGG pathway
305 containing more than 20 EPOCs and with a greater aELW than Asgard (Figure 4; see supplementary
306 Figure 6A for alternate cutoffs). Specifically, oxidative phosphorylation shows a strong signal for
307 Alphaproteobacteria (see supplementary Figure 3 for OXPHOS complexes) while Myxococcota (formerly,
308 Deltaproteobacteria) showed a consistent associations with metabolic pathways involved in nicotinate,
309 nucleotide, and nucleoside-sugar modification and metabolism. However, these nucleotide-related
310 associations appear to be limited to phosphatases and phosphoribosyltransferases acting on nucleoside-
311 sugars, specifically 5' and 3' nucleotidases and are as such much more limited in functional scope than
312 pathways of Alphaproteobacterial associations. Previous studies highlighted a link between
313 Myxococcota and steroid biosynthesis in Eukarya [60]. While we observed a clear, if narrow (6 EPOCs),
314 association with Myxococcota for diterpenoid biosynthesis, downstream of the mevalonate pathway
315 and upstream of steroid synthesis, our criteria on coverage and taxonomic scope limited the
316 representation of downstream steroid pathways in the core data (supplementary Figure 5 and 6A).
317

318 Although few specific associations were observed with diverse bacteria on the level of pathways, the
319 possibility remained that individual proteins with key functions could originate from such taxa. For this
320 analysis, we considered a stricter subset of the core EPOCs, requiring a eukaryotic outgroup containing
321 at least 15 taxonomic clades, to ensure functional centrality, and prokaryotic sister taxa with at least 20
322 sequences and an ELW > 0.7 to ensure strong association. Of the 130 unique EPOCs meeting these
323 criteria, 33 were found to be associated with diverse bacterial lineages, mostly Actinomycetota and FCB
324 group bacteria(supplementary Figure 6B). The remaining 97 EPOCs were found to be Asgard-derived.
325 The 33 EPOCs of prominent bacterial origin covered a wide range of cellular functions from major
326 facilitator superfamily transporters to lipases, to components of core sugar metabolism and cardiolipin
327 synthesis. Once again, we observed no clear functional trend for these bacterial contributions. Thus,
328 although single protein associations were individually significant, no pathways appeared to have been
329 derived from any single bacterial taxon. Instead, we interpret these diffuse associations with diverse

330 bacteria as indications of multiple bacterial contributions, each with a limited functional scope, likely
331 originating from isolated HGT events rather than additional endosymbioses.
332



333
334 **Figure 4. Strongest bacterial contributions to eukaryotic pathways compared to the contributions of**
335 **Asgard archaea.** Each prokaryotic taxon is represented by the highest aELW of any KEGG BRITE pathway
336 with at least 20 EPOCs, relative to that of Asgard for the same pathway (background, pale blue). The
337 strongest prokaryotic association relative to Asgard was that of Myxococcota with Nicotinate
338 metabolism followed by Alphaproteobacteria with oxidative phosphorylation.
339

340 ## Ancestral stem lengths and the amount of evolution pre- and post-LECA

341 We attempted to use tree stem length analysis to infer the timing of acquisition of core eukaryotic
342 genes, as originally implemented by Pittis and Gabaldon [35], with the mitochondrion as a test case. As the
343 time elapsed post-LECA is the same for all genes, it has been proposed that the relative stem length in
344 gene trees is proportional to the time elapsed since the acquisition of the gene from its prokaryotic
345 donor to the LECA. Using this approach, Pittis and Gabaldon found that proteins of Alphaproteobacterial
346 descent had significantly shorter stems than proteins of archaeal descent, consistent with a
347 mitochondria-late scenario for eukaryogenesis [33-35]. The raw stem length is defined as the distance from
348 the Last Common Ancestor (LCA) of Eukarya and the most likely prokaryotic sister phylum, the First
349 Eukaryotic Common Ancestor (FECA)[48,61], to the LECA. To account for differences in evolutionary rates,
350 this raw stem length is normalized by dividing by the median eukaryotic branch length, measured from
351 the LECA (Figure 5A).
352

353 Across the entire set of eukaryotic stem lengths for our dataset (5,850 stems), we observed a wide
354 distribution with a sharp maximum close to 0.05, highly reminiscent of previous findings [35] (Figure 5B

355 5C). In our analysis, Alphaproteobacterial stems shorter than 0.3 (70% of all stems) were significantly
356 longer than those of Asgard origin, but Alphaproteobacterial stems longer than 0.35 were instead
357 shorter than Asgard on average, showing a non-obvious yet reproducible (supplementary figure 7B) and
358 statistically significant relationship (Figure 5B). Stems of Cyanobacterial origin followed a similar but
359 weaker trend. Thus, although the observation of globally shorter stem lengths for genes of
360 alphaproteobacterial descent applied only to stems longer than 0.35, we effectively recapitulate the
361 findings of Pittis and Gabaldon, who primarily analyzed stems within that range. When we limited our
362 analysis to ribosomal proteins and components of oxidative phosphorylation complexes, that is, genes
363 of well-known and easily identifiable Archaeal and Alphaproteobacterial ancestries, respectively, the
364 above trends were further accentuated (Figure 5C). Contrary to the notion of stem lengths as a measure
365 of the acquisition timing, we observed variances of more than two orders of magnitude in normalized
366 stem lengths for the oxidative phosphorylation complexes that are assumed to have been acquired
367 simultaneously during mitochondrial symbiogenesis (Figure 5B). Taken together, these findings suggest
368 that even normalized stem lengths are determined, primarily, by factors other than the gene acquisition
369 time, at least, across the long evolutionary timescales analyzed here.
370
371 Comparison of the stem length distributions across functional classes of genes (Figure 5A) prompted a
372 possible explanation for these observations. Shorter than average stems were observed for genetic
373 information processing, belying the expectation of these genes being the longest-residing genes in the
374 nascent eukaryotic lineage. We suggest that another major factor determining the relative stem length
375 of a gene might be the amount of evolution post-acquisition, necessary to adjust an acquired gene to a
376 new biological environment. The genes inherited from the Asgard ancestor would have been pre-
377 adapted to the cellular environment of the evolving protoeukaryotes, whereas genes acquired from
378 radically different bacterial sources had to substantially adapt post-acquisition, increasing the apparent
379 lengths of their pre-LECA stems, which is at least partially supported by our general observation of long
380 stems from diverse bacteria (supplementary Figure 7C).
381

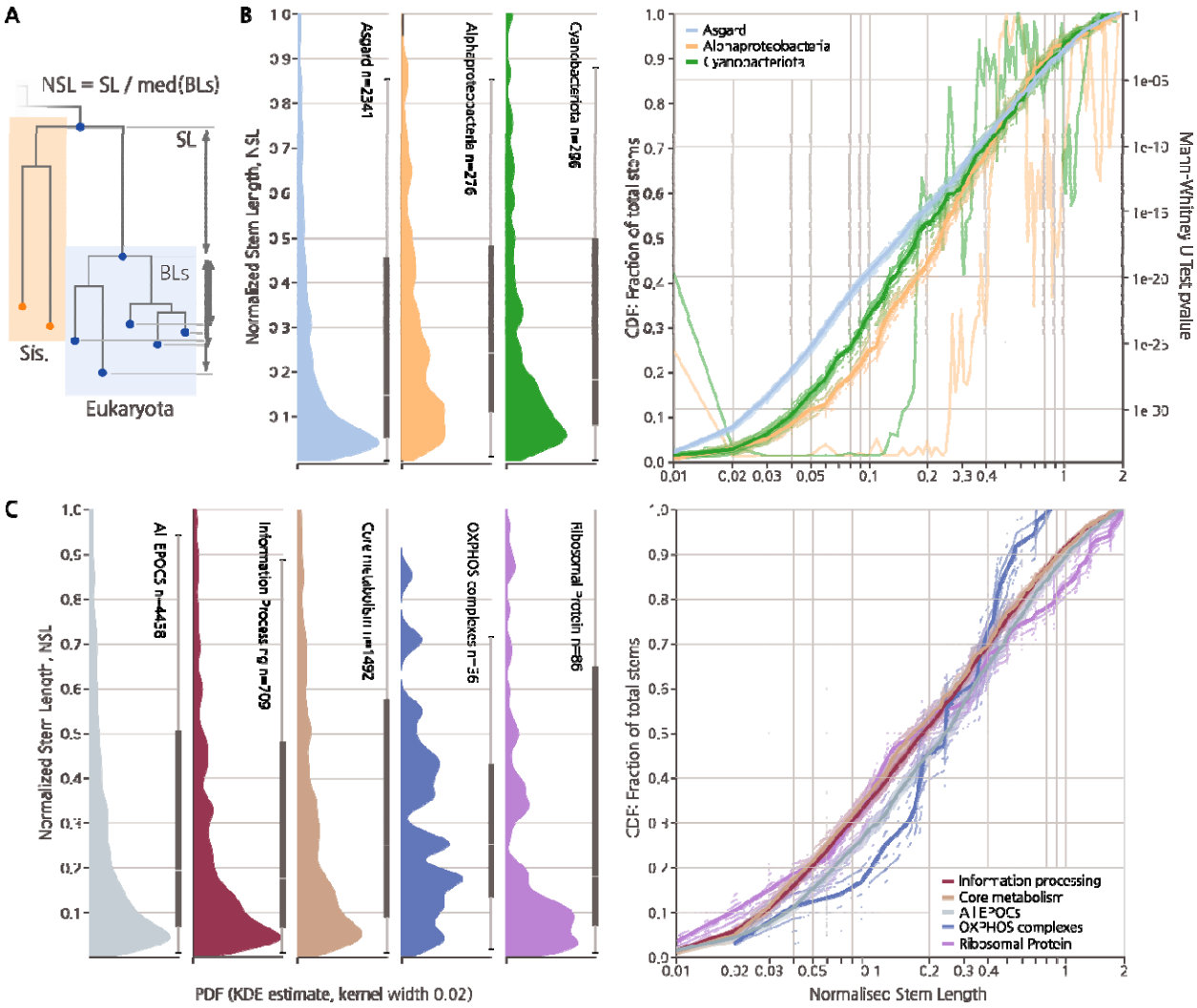


Figure 5. Ancestral stem length distributions for different groups of prokaryotes contributing to eukaryogenesis and different functional classes of genes. A) Schematic of stem length calculation. Stem Lengths (SL) were calculated as the lengths of branches from the shared root of Eukarya and the inferred ancestral prokaryotic sister phylum to the root of the Eukarya clade. Normalized Stem Lengths (NSL) shown in A and B were calculated by dividing SL by median Branch Lengths from the root of Eukarya to individual Eukarya leaves (BLs). **B)** Left, Probability Density Function (PDF) of normalized stem lengths taken from samples of individual core eukaryotic genes associated with prokaryotic taxa, obtained through KDE. Thin bar indicates 5th percentile range, thick bar includes 25th percentile ranges with median indicated. Right, fractional Cumulative Distribution Functions (CDF) of normalized stem lengths together with 20 of 200 bootstrap replicates from EPOCs within each category. Second Y loglinear axis indicates value of the 2-sided Mann Whitney U test of bootstrap CDF distributions per stem length between Asgards and either Cyano- or Alphaproteobacteria. **C)** As in B with genes grouped by BRITE A Metabolism (KO:09100), Genetic Information Processing (KO:09120), Oxidative phosphorylation (map00190 excluding KOGs for V-type ATPases).

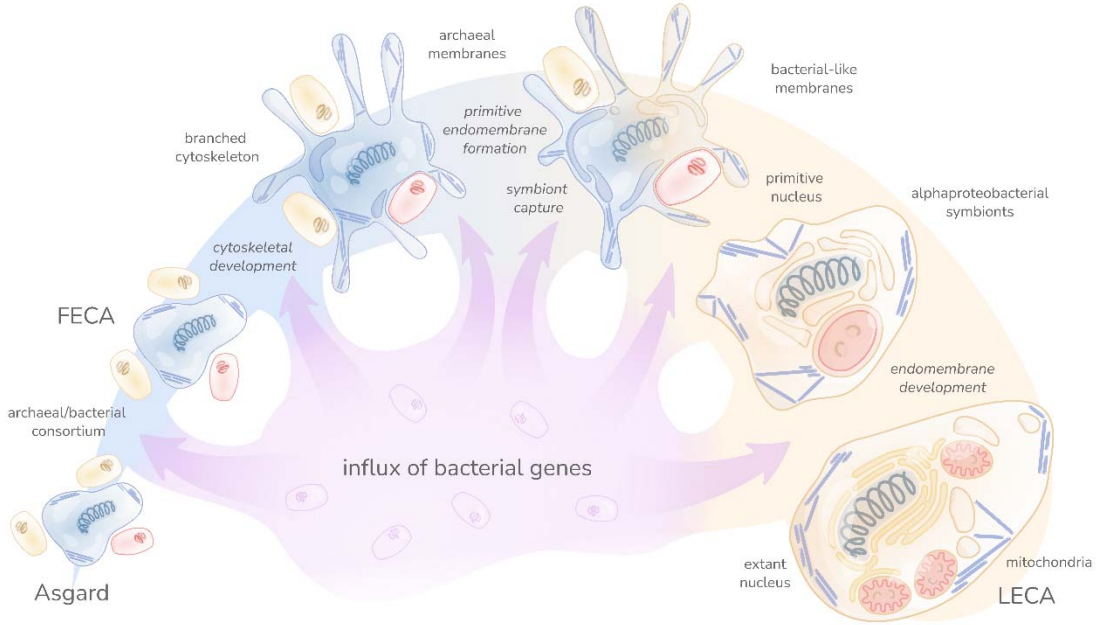
397 ## Discussion

398 Reconstruction of eukaryogenesis is a moving target, as strikingly demonstrated by the discovery and
399 expanding exploration of Asgard archaea which revolutionized the field [2-7]. Nevertheless, recent
400 progress in prokaryote and eukaryote genome sequencing and the advances of metagenomics have
401 created unprecedented opportunities for understanding the phylogenomic origins of eukaryotes. In this
402 work, we took advantage of this expanded genome collection, coupling it with maximally sensitive HMM
403 profile-profile searches, and performed a comprehensive phylogenetic analysis of core eukaryotic genes
404 tentatively mapped to the LECA, within a statistical framework focused on testing evolutionary
405 hypotheses using constraint trees.
406
407 Many previous studies, especially early ones, demonstrated the apparent chimeric origins of the core
408 eukaryotic gene set, with a bacterial contribution quantitatively exceeding the archaeal one, the latter
409 being largely limited to information processing [17-19,62]. The discovery and exploration of Asgard archaea
410 partially changed this notion by demonstrating that many genes involved in various cellular processes
411 and systems, such as membrane remodeling and cytoskeleton, were of apparent Asgard origin [5,6,8].
412 Indeed, the presence of an elaborate cytoskeleton in Asgard archaea has been subsequently
413 experimentally demonstrated [24]. The present work substantially extends the dominance of Asgard in the
414 ancestry of the eukaryotes by demonstrating a likely Asgard origin for the majority of genes traced to
415 the LECA, and a major Asgard association for nearly all functional systems and pathways of the
416 eukaryotic cell. In only a handful of metabolic pathways did the detected bacterial contribution exceed
417 that of Asgard archaea. The most conspicuous of the bacterial contributions, not unexpectedly, comes
418 from Alphaproteobacteria, and accounts for the core components of the electron transport chain, iron-
419 sulfur cluster metabolism, and the mitochondrial translation system. However, although prominent and
420 functionally coherent, the Alphaproteobacterial contribution is relatively small, not comparable to that
421 of Asgard, and notably, almost entirely restricted to mitochondrial functions. The contributions of other
422 bacterial phyla, although cumulatively substantial, and in select cases strongly supported for individual
423 EPOCs, failed to show any consistent functional trends, with the exception of notable but limited
424 associations between Myxococcota and the synthesis of nucleotides, nucleoside-sugars and steroid
425 precursors.
426
427 These findings are best compatible with a scenario of eukaryogenesis in which the Asgard FECA (the
428 base node of the lineage leading from Asgard to the LECA) already possessed many signature features of
429 eukaryotic cells, including the cytoskeleton and the endomembrane system as posited in some previous
430 models [26,63]. Under this scenario, mitochondrial acquisition was a relatively late event that made a
431 limited, but functionally crucial, contribution to the gene composition of the emerging eukaryote (Figure
432 6). Other bacterial contributions appear to be piecemeal, possibly, stemming from a continuous but
433 non-specific capture of genes from diverse bacteria during the evolutionary path from the ancestral
434 Asgard archaeon to the LECA (Figure 6). The syntrophic lifestyle of many Asgard archaea [32] is conducive
435 to extensive gene influx from bacteria, enhancing the general trend of mosaic gene composition of
436 prokaryotic genomes [71,72]. Much of the bacterial gene acquisition likely preceded the emergence of the
437 proto-eukaryotic lineage and more occurred along the path from the FECA to the LECA (Figure 6). Thus,

some of the core eukaryotic genes, although proximally inherited from the Asgard ancestor, might ultimately be of bacterial origin, having been carried via Asgards into the FECA. As such, we obtained no indication of another symbiotic event contributing to eukaryogenesis.

How eukaryotes acquired the bacterial-type membranes remains an enigma. Whereas the eukaryotic isoprenoid synthesis appears to have been derived from Asgard lipid synthesis, the ancestry of the enzymes involved in membrane lipid biosynthesis appears complex and possibly mosaic, remaining unresolved in the present analysis. Nevertheless, the Asgard affinity of adjoining systems and pathways, such as the synthesis of GPI anchors, sphingolipids, and isoprenoids, suggests that a substantial component of the lipid biosynthesis and modification capacity of eukaryotes persisted from the Asgard ancestor. If an Asgard archaeal cell was the principal scaffold on which the eukaryotic cell evolved, replacement of the archaeal membrane with a bacterial one must have occurred at an early stage of eukaryogenesis, progressing through a mixed membrane stage. Viable bacteria with a mixed bacterial-archaeal membrane have been experimentally engineered, albeit typically with a substantial fitness decrease [64,65,79], demonstrating the possibility of such a scenario. Membrane replacement might have even predated mitochondrial endosymbiosis, following early capture of the necessary bacterial enzymes by the Asgard archaeon (Figure 6).

A major issue for understanding eukaryogenesis is the origin of the nucleus, the ubiquitous, eponymous organelle of eukaryotic cells [66,67]. Our analysis supports the near complete origin of information processing systems located in the nucleus from Asgard ancestors. However, given that the nuclear envelope membrane is an extension of the ER and as such is of the bacterial type [68], whereas the nuclear pore consists largely of repetitive proteins of uncertain provenance [69], the origin of the nucleus as a membrane-bounded compartment remains an unresolved question that has to be addressed in conjunction with the membrane replacement problem.



465 **Figure 6. Hypothetical scenario of eukaryogenesis.** The figure shows the gene flow along the succession
466 of events leading from a simple archaeal ancestor through the more complex Asgard intermediate
467 capturing the proto-mitochondrial Alphaproteobacterial endosymbiont to the LECA. The wide colored
468 shape denotes the dominant trend of the origin of eukaryotic genes from Asgard ancestors, with light
469 blue portions corresponding to archaeal evolution and the light brown portion to the evolution of
470 protoeukaryotes starting with the Alphaproteobacterial endosymbiosis. The purple arrows denote the
471 smaller, piecemeal contributions of gene influx from various bacteria that is posited to have been
472 occurring at all stages of the depicted evolutionary scenario.
473
474 The conclusions of this work are subject to caveats stemming, above all, from the biased and still limited
475 sampling of sequenced archaeal and bacterial genomes. In particular, almost all currently available
476 genomes of Asgard archaea are incomplete metagenomic assemblies, with the exception of only a few
477 closed circular genomes [22-24,31,70] which inevitably leads to uncertainty in the estimation of the Asgard
478 contribution to eukaryogenesis. The same pertains to Alphaproteobacteria, both because of the
479 sequencing bias whereby many of the sequenced alphaproteobacterial genomes come from symbionts
480 and parasites with reduced gene complements, and because the ancestor of the mitochondria
481 apparently belongs outside of the currently known diversity of Alphaproteobacteria [71]. Other bacterial
482 phyla are even more severely under-sampled, potentially, leading to underestimates of their
483 contributions to the evolution of eukaryotes. For example, given the limited genomic information, it is
484 difficult to rule out that we capture only part of a major signal from Myxococcota, proposed as one of
485 the partners in the syntrophic scenario of eukaryogenesis [15,22]. Our pangenome-based approach to
486 exclude late HGT events from ancestral inference potentially could overlook possible contributions from
487 narrow or severely under-sampled clades and might be insufficient to compensate for extensive and
488 continuous horizontal transfer across diverse eukaryotic taxa. Additionally, if future work indicates a
489 substantial mismatch between the representations of eukaryotic biology in KEGG and in a reconstructed
490 gene complement of the LECA, our results should be revisited. Expanding sampling of prokaryotic
491 genome diversity and further technical improvements in horizontal transfer detection might in the
492 future lead to a substantial modification or even complete revision of the model of eukaryogenesis. It
493 nevertheless appears that the dominant contribution of Asgard archaea, the main finding of this work, is
494 here to stay.
495

496 # Methods
497
498 ## Database curation
499 For prokaryotes, 75 million sequences were curated from 47545 fully sequenced prokaryotic genomes
500 obtained from the NCBI GenBank (https://ftp.ncbi.nlm.nih.gov/genomes/) in November 2023
501 (prok2311) and supplemented with 441150 proteins sequences from 63 Asgard metagenomes [4,5]
502 (prok2311As), selected to represent all clades in these two publications where available. Protein
503 sequences were either directly taken from GenBank annotations, or produced by Prodigal v2.6.3 [72]
504 trained on the set of 12 complete or chromosome-level assembled Asgard genomes. To only include
505 those sequences which are widely present within prokaryotic families, and to minimize the possibility of
506 post-LECA horizontal transfer from eukaryotes, soft-core pangenomes were constructed for each of our
507 26 curated taxonomic groups, mostly conforming to the NCBI taxonomy rank "class" (see below and
508 supplementary Data 1). To construct the soft-core pangenomes, all sequences from each taxonomic
509 group were clustered individually using mmseqs2 [75] "mmseqs cluster -s 7 -c 0.9 --cov-mode 0" [75]. All
510 clusters which did not contain sequences from at least 50% of all bacterial species (here defined by
511 lowest rank taxonomic ID within the NCBI taxonomy), or 20% of eukaryotic sequences (see below), were
512 rejected from the database. The effects of alternative pangenome definitions, such as 10, 24, 50, and
513 67% proteins per class, are shown in supplementary figure 1D.
514
515 The eukaryotic database was constructed starting from a curated set of 72 genomes available on the
516 NCBI GenBank. In order to provide a more accurate representation of eukaryotic diversity, this dataset
517 was supplemented with EukProt v3 [73], a curated database aiming to provide a sparse representation of
518 eukaryotic biology, to create Euk72Ep containing 30.3 million sequences. As sequences included in
519 EukProt v3 come from a diverse range of sources, some known to contain contaminant prokaryotic
520 sequences, the database was screened for prokaryotic contamination. This screen was carried out by
521 first constructing candidate prokaryotic and eukaryotic HMM databases as described below and
522 querying both databases against the eukaryotic sequence database using hhblits [40]. All eukaryotic
523 sequences with a top hit alignment score from a prokaryotic HMM were removed from the database
524 prior to initial clustering.
525
526 ## Curation of taxonomic labels
527 In order to provide a compromise between taxonomic specificity and accuracy of clade assignment, we
528 curated a set of taxonomic labels for both Euk72Ep and Prok2311As to act as representatives. To
529 construct this set, we started with manually assigning all species in EukProt to their closest relative in
530 the NCBI taxonomy. Then, each species from Prok2311As and Euk72Ep was assigned a taxonomic label
531 corresponding to their "Class" rank (for example, Alphaproteobacteria, Thermococci, Mammalia) based
532 on the NCBI Taxonomy of November 2023 using tools from the ete3 3.1.3. toolkit [74] as well as custom
533 scripts. For species without a corresponding "Class" rank, the closest relevant rank was manually
534 assigned. This procedure yielded an initial list of 317 unique taxonomic class identifiers. As the rank
535 "Class" does not evenly partition biological diversity, identifiers were further manually mapped to a
536 curated set of 45 eukaryotic and 26 prokaryotic taxonomic superclasses, each present in the NCBI

537 taxonomy (for example, Metazoa, Streptophyta, TACK archaea; see supplementary Data 1. Due to the
538 partially unresolved taxonomic relationship within Asgard archaea, all candidate Asgard sequences from
539 Prok2311 or from references 6,7 were assigned the class label "Asgard". Following the recent
540 reclassification of Deltaproteobacteria into Myxococcota, Desulfobacterota, Bdellovibrionota and
541 SAR324 species which could not be mapped into either of these clades using existing taxonomic
542 annotation were classified as Deltaproteobacteria (0.025% of total data).
543

544 **Sequence clustering and profile database generation**

545 To transform Euk72Ep and Prok2311As into profile databases suitable for sensitive HMM-HMM
546 searches, we implemented an unsupervised, cascaded sequence-profile clustering pipeline using the
547 tools available in the mmseqs2 software suite [75]. In brief, each sequence database was initially clustered
548 at 90% sequence identity and 80% pairwise coverage using "mmseqs linclust" and cluster
549 representatives were chosen using "mmseqs result2repseq". This procedure provides a non-redundant
550 set of 5.9 million prokaryotic and 12.7 million eukaryotic sequences for further analysis. The non-
551 redundant sequences are further collapsed into a set of initial profiles and consensus sequence pairs
552 using "mmseqs cluster", "mmseqs result2profile" and "mmseqs profile2consensus". All consensus
553 sequences are queried against their profiles using "mmseqs search" and results clustered using
554 "mmseqs clust". Clusters are mapped to their original non-redundant sequences using "mmseqs
555 mergeclusters" and profiles and consensus sequences are constructed once again. In order to grow
556 clusters based on their sequence-profile alignments, this procedure of "mmseqs profile2consensus -
557 mmseqs search - mmseqs clust - mmseqs result2profile" is iterated until cluster sizes converge. An 80%
558 pairwise coverage is maintained throughout all steps of the cascaded clustering protocol to avoid
559 homologous overextension. The final clustered databases contain 14.1 million and 91.000 clusters for
560 Euk72Ep and Prok2311As, respectively.
561

562 To transform the unsupervised clusters into a set of HMM profiles, we employed a mixed MSA
563 construction strategy coupled with automatic data reduction to create accurate alignments for a wide
564 diversity of cluster sizes and sequence diversities. To construct HMMs, all non-singleton sequence
565 clusters are first aligned using FAMSA 2.0.1. [76] to produce an initial alignment. Using the initial
566 alignment, we then reduce the sequence space by calculating an approximate tree using FastTree 2.1.1.
567 using "FastTree -gamma" [77], remove long branch outliers as described in "EPOC tree construction and
568 processing", and iteratively removing the closest pairwise leaves using ete3 [74], keeping the leaf closest
569 to the root, until the desired sequence set size is reached. For profile generation, we retain no more
570 than 300 sequences for both Prok2311As and Euk72Ep. This reduced sequence set is then realigned with
571 muscle5 5.1.linux64 [43] using "muscle -diversified" with 5 replicates, and the maximum column
572 confidence alignment is extracted using "muscle -maxcc". This alignment is then trimmed to keep
573 columns with more than 0.2 bits of Shannon information defined as the difference of log2(20) and the
574 Shannon entropy of the column amino acid distribution. This process of data
575 reduction/alignment/trimming is referred to as "prune-and-align" and employed later for downstream
576 analysis. Trimmed alignments are turned into an HMM database using HH-suite with "hhconsensus -M

577 50", hhmake and "cstranslate -f -x 0.3 -c 4 -l a3m". The final HH-suite databases contained 480.000 and
578 1.6 million profiles for prokaryotes and eukaryotes, respectively.
579

580 ## Cluster annealing

581 As the initial mmseqs2 based clustering was carried out in an unsupervised manner, we considered it
582 necessary to ensure that the results were not contingent on this unsupervised partitioning of the data.
583 To this end, a further cluster annealing step was developed which aims to redefine clusters based on
584 clades well separated monophyletic clades on phylogenetic trees, rather than by sequence similarity
585 alone. Starting with the existing clustered databases Euk72Ep and Prok2311As, HMM profiles were
586 calculated for each non-singleton cluster and all vs all HMM-HMM searches were performed using
587 Hhblits 3.3.0 (HHblits -n 1 -p 80), retaining hits with at least 80% HHblits probability and 50% pairwise
588 profile coverage. The resulting hits were then clustered using greedy set clustering, as described in
589 mmseqs2, implemented in python. These superclusters represent the limit of detection using out
590 sequence-based methods, but might involve overclustering, possibly merging multiple similar gene
591 families into one set. To mitigate the potential effect of such overclustering, cluster partitioning was
592 performed based on phylogenetic tree analysis. To this end, the sequence space of the superclusters
593 was first reduced using the "prune-and-align" approach described above, after which a master tree was
594 constructed using FastTree -gamma. Given the resulting set of trees, we sought to identify well
595 separated monophyletic sets of leaves, the presence of which would warrant partitioning of the
596 respective superclusters. To perform this tree partitioning in an unsupervised manner, the all-vs-all
597 pairwise leaf distances were calculated as a square matrix and embedded into 2 dimensions using UMAP
598 version 0.5.3 (n_neighbors=50, min_dist=0.3, distance='euclidean'). This embedding was clustered using
599 HDBSCAN as implemented in sklearn version 1.3.0 with default parameters. We explicitly allow single
600 cluster partitions to ensure that trees with uniform distributions of pairwise branch lengths remained as
601 single clusters. Mapping clusters back to the phylogenetic trees showed that UMAP embedded
602 HDBSCAN clusters were well separated in the original trees (Supplementary figure 8). These annealed
603 clades were used in downstream processing, representing a stricter definition of EPOCS than the initial
604 cascaded mmseqs2 clusters. The annealing process redistributed sequences from the 1.6 million
605 eukaryotic clusters into 1.5 million annealed clusters and from 460.000 prokaryotic sequences into
606 280.000 annealed clusters, highlighting the already dense representation of initial mmseqs2 clusters.
607 Because the embeddings were done per tree, the partitioning into subsets is independent of branch
608 lengths and merely relies on the internal distribution of data. Thus, well-formed subclades could be
609 identified in sequence sets independent on the internal diversity, an essential feature when applying
610 this procedure.
611

612 ## Search for prokaryotic homologs of eukaryotic proteins

613 Given that clade-specific protein families were not of direct relevance to this study, we reduced the
614 search space by excluding HMM profiles from eukaryotic clusters with less than 10 sequences and a
615 lowest common ancestor with a taxonomic rank below "Superkingdom" as calculated by "mmseqs lca".
616 This requires a sequence cluster to contain at least two sequences from different kingdoms, as defined
617 by the NCBI taxonomy (ex. Amoebozoa and Haptista, or Viridiplanta and Opistokonta), retaining 142,000

618 profiles of conserved eukaryotic proteins. To associate eukaryotic sequence clusters with homologous
619 prokaryotic clusters, we queried the Euk72Ep HMM database against all Prok2311As HMM profiles using
620 a single iteration of HHBlits 3.3.0 as "HHblits -n 1 -p 80" and retaining the hits with at least 80% HHblits
621 probability and 80% pairwise profile length. These filters resulted in a total of 20,700 accepted
622 eukaryotic query profiles targeting a total of 8,300 unique prokaryotic profiles. These clusters contained
623 a total of 5.7 million and 1.9 million sequences.
624

625 **EPOC multiple sequence alignment construction**

626 Unless otherwise stated, all subsequent steps were carried out using custom python 3.9 scripts, with
627 taxonomy and tree parsing carried out using ete3 [74]. All sequences from a eukaryotic cluster with at
628 least 10 sequences and a lowest common ancestor at the rank of "Superkingdom" (NCBI taxonomy),
629 together with all members of the homologous prokaryotic clusters identified in the HMM-HMM search
630 with at least 80% probability and pairwise sequence coverage, forms an EPOC. As EPOCs varied in size by
631 more than 5 orders of magnitude (10 to 100,000 sequences), robust data subsampling is necessary for
632 accurate downstream MSA generation and tree construction. We employed a variant of the "prune-and-
633 align" strategy described above for profile generation taking into account the taxonomic distribution of
634 the constituent sequences. Rather than iteratively pruning the closest leaves pairwise, we first label
635 each leaf with its corresponding taxonomic label and only prune monophyletic groups, maintaining a
636 count of all pruned leaves per clade. If the procedure fails to reach the target leaf number before
637 converging, all leaves are ordered by the number of pruned relatives and leaves with the lowest number
638 of relatives deleted (retaining the largest monophyletic groups) until the target size is reached. The
639 result approximates a maximum diversity representation, preferentially pruning isolated singleton
640 clades. Using this modified protocol, eukaryotic and prokaryotic sequences are cropped to a maximum
641 of 30 and 70 sequences, respectively. Each EPOC therefore consists of a maximum of 100 representative
642 sequences. All EPOCs are aligned using "muscle -diversified" with 5 replicates, and the maximum
643 confidence alignment is extracted using the --maxcc argument. This alignment is then trimmed to keep
644 columns with more than 0.15 bits of Shannon information content for tree reconstruction.
645

646 **EPOC tree construction and processing**

647 From the pruned and trimmed alignment, we created a maximum likelihood tree using IQtree2 2.3.5.
648 "IQtree2 -B 1000 -bnni -m MFP -mset LG,Q.pfam --cmin 4 --cmax 12" with model parameters estimated
649 by model finder plus [44]. Candidate models and rate category search ranges were selected based on a
650 test set of 200 randomly chosen EPOCs from the filtered core set analyzed with full model parameter
651 evaluation by Model finder [44] and 10 tree replicates. The LG and Q.pfam models and rate category
652 ranges were chosen as they consistently produced the highest ranking log-likelihoods for a random
653 sample of 200 EPOCs run with full model estimation in IQtree.
654

655 As the master trees are built from sequences obtained through sensitive cascaded sequence-profile
656 clustering we would in rare cases observe long-stemmed clade outliers, in addition to more common
657 long stem leaf outliers. In order to assess and remove such erroneous data, we first estimated the log-
658 normal distribution of all stem lengths using scipy.stats 1.11.1 and excluded any stems outside the 0-

659 99.5 % Probability Point Function interval. This simultaneously removes long terminal leaf branches as
660 well as highly diverged clades. In rare cases where this pruning would remove either the entire
661 eukaryotic clade or more than 30 % of all leaves, the EPOC was discarded (387 EPOCs). The resulting
662 trees were then re-rooted using a weighted midpoint approach so that the sum of all branch stem
663 lengths on either side of the root was equal. For clarity, no trees in our current study have been
664 "phylogenetically rooted" in the biological sense. All references to the "root" refer to the origin of the
665 data structure, not the biological concept.
666
667 To assign taxonomic clades, all leaf nodes were assigned taxonomic labels from our curated set and
668 monophyletic clades were collapsed, taking as the representative the lowest branching leaf. We found
669 that assigning clades purely based on a monophyletic definition of taxonomic purity severely hampered
670 the analysis as leaves can be taxonomically incongruous with their neighbors. This effect most likely
671 stems from cases of locally erroneous tree topology or local HGT, but severely complicates the
672 downstream analysis. We therefore adopted the notion of a "soft LCA" representing the root of a close-
673 to-monophyletic clade. To greedily rank all the best "soft LCAs", for each taxonomic label, the tree was
674 traversed from each monophyletic clade root to the tree root, calculating a score for each internal node:
675
676 (number of leaves with label X in clade / clade size) * (number of leaves with label X in clade / total
677 number of label X)
678
679 This metric balances taxonomic "purity" and "scope" for each possible clade of label X. All such nodes
680 are then ordered by this score. Because this score is sensitive to the total number of taxonomic labels in
681 an EPOC, it was not used as a filtering criterion at any point in the analysis, yet is present in the
682 metadata, mentioned here for completeness. To avoid overinterpreting small clades, clades were
683 considered valid if they represented at least 3 sequences with clade purity > 0.8 for prokaryotes, and at
684 least 5 sequences with purity > 0.8 for eukaryotes. Trees that fail to identify either any eukaryotic or any
685 prokaryotic clades under these constraints, primarily derived from small EPOCs, are discarded (411
686 EPOCs). Trees with more than 3 valid eukaryotic clades were considered to show high paraphyly and
687 likewise discarded (781 EPOCs).
688

689 **Evolutionary hypothesis testing using constraint trees**

690 In order to assess the relative probabilities of each sampled prokaryotic clades to represent the
691 eukaryotic sister clade, we employed evolutionary hypothesis testing using constraint trees as
692 implemented in IQtree2 [44]. For each EPOC and for each detected eukaryotic clade, we generated a set of
693 constraint trees, one for each possible prokaryotic sister. Each constraint tree had three defined clades
694 enforced throughout tree calculations: 1) the eukaryotic group as defined by all sequences below the
695 eukaryotic LCA, 2) one of the prokaryotic sister groups and 3) all other prokaryotic leaves. We then
696 construct a set of local trees from slices of the original MSA using IQtree2, forcing the constraint tree
697 topology using "iqtree2 -g", using the same evolutionary model as the original, unconstrained "master
698 tree". The set of constrained trees is then ranked using "iqtree2 -z" calculating the relative model
699 assignment confidence for each of the constrained topologies and resulting test metrics are saved for

700 downstream analysis. As the number of trees to evaluate scales by the number of eukaryotic clades
701 multiplied by the number of prokaryotic clades, we limit the sampling to a maximum of 3 clades per tree
702 for eukaryotes, discarding rare cases of high eukaryotic paraphyly, and only consider the 12 closest
703 prokaryotic clades to each individual eukaryotic clade, based on their topological distance calculated as
704 the number of non-root tree bifurcations from one node to another (supplementary Figure 2). The
705 procedure results in an Expected Likelihood Weight (ELW)[41] for each of the evaluated model trees, here
706 taken as being analogous to model selection confidence that an evaluated prokaryotic clade is the most
707 significant sister to a eukaryotic clade.
708

709 **EPOC annotation**

710 To functionally annotate protein families within EPOCs, we generated profiles based on protein
711 sequences in KEGG release 110 [42]. For each KOG, we extracted Kegg metadata including KEGG pathways
712 and BRITE labels via KEGG's API. At the time of parsing (January 2024), KEGG contained 26,695 KOGs.
713 Due to KEGG API constraints, protein sequences in KEGG were extracted from Uniprot, first by mapping
714 KEGG proteins to Uniprot IDs via Uniprot's ID mapping service (uniprot.org/id-mapping), then by
715 extracting the sequences for these Uniprot IDs corresponding to each KOG [78].
716

717 For each KOG, we generated HMM profiles using the aforementioned prune-and-align pipeline. To
718 increase specificity of labeling, we partitioned those KOGs which encompass both prokaryotic and
719 eukaryotic sequences into separate taxonomic groups and sequence alignments were generated
720 separately for prokaryotes and eukaryotes. Each eukaryotic HMM profile forming the basis for a EPOC
721 was queried against our KEGG profile database using "HHblits -n 3 -p 80". Results were subsequently
722 filtered to a pairwise profile coverage of 0.7 using custom scripts. As a result, 12,600 of the 13,500
723 EPOCs could be annotated by at least one KOG profile at 80% probability and 70% pairwise coverage.
724 The highest probability hits were used for annotation.
725
726

727 **Data filtering and removal of cases of possible late horizontal gene transfer**

728 Unless otherwise stated, all data is interpreted from the core set of EPOCs meeting the following
729 criteria: 1) the eukaryotic constituent cluster profile identifies at least one target profile in our KEGG
730 profile database at a probability of 80% and 70% coverage, 2) the eukaryotic clades include more than 5
731 distinct taxonomic labels, at least one from Amorphea and one from Diaphoretickes, encompassing
732 LECA, 3) only prokaryotic sister clades with 0.4 < ELW < 0.99 were considered reliable and were included
733 in the analysis. We note the independence on aELW values on the choice og number of wukaryotic
734 clades as showin in Supplementary Figure 7A. Cases of ELW >= 0.99 were not considered as these are
735 found to be disproportionately eukaryotic-like, often branching from within the eukaryotic clade itself,
736 and thus likely derived from late horizontal transfer post LECA. Due to the presence of such high ELW
737 values within Alphaproteobacterial associations with Oxidative phosphorylation, in this case, as an
738 exception, ELW values of 1 were included.

Taxonomy remapping and EPOC construction under GTDB taxonomy

To ensure that the results were robust to taxonomy, the prok2311 database was reformatted under the GTDB taxonomy (version 220) at the level of "phyla" [47]. For taxonomic mapping, first, genomes present in both prok2311 and GTDB were directly annotated with the GTDB taxonomy. For the remaining genomes, taxonomy was assigned through consensus voting by using the set of GTDB marker genes, searched against our data using "mmseqs search -s 7 -c 0.5 -e 1e-10". A phyla level taxonomic label was assigned to each genome based on the largest number of total hits. Genomes without significant hits to GTDB marker genes were discarded for the purpose of this analysis (0.92% of all data). This procedure mapped the prok2311 data to 96 GTDB phyla which were used as taxonomic assignment for EPOC calculations as described above. All additional Asgard data taken from Ref 8 were directly assigned as Asgardarchaeota.

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

758 **Data availability**
759 All final non-redundant databases with taxonomy annotation, clustering data, HMM profiles, EPOCs data
760 and final parsed and tabulated results and metadata is available at Zenodo
761 https://zenodo.org/records/15048010
762
763 **Code availability**
764 All code and scripts used to generate the final parsed data is available on Github at:
765 https://github.com/VictorTobiasson/eukgen
766
767 **Supplementary Data File 1**

768 Excel file with four sheets: 1) List of revised taxonomy mapping species without well curated taxonomy
769 to closest existing taxonomic ranks. 2) List of revised taxonomy merging sparsely populated classes. 3)
770 Manually curated list of genes involved in Fe-S assembly. 4) Tabular data of all tables included in figures.
771
772 **Author contributions**
773 E.V.K. initiated the study; V.T. and Y.I.W. developed the pipeline for data analysis; V.T. and J.L. collected
774 the data; V.T., J.L., Y.I.W and E.V.K. performed data analysis; V.T. and E.V.K wrote the manuscript which
775 was read, edited and approved by all authors.
776
777 **Acknowledgements**
778 The authors thank Toni Gabaldon and Koonin Group members for helpful discussions. The authors'
779 research is supported by the Intramural Research Program of the National Institutes of Health of the
780 USA (National Library of Medicine).
781

782 **Competing interests**

783 The authors declare no competing interests.

784

785

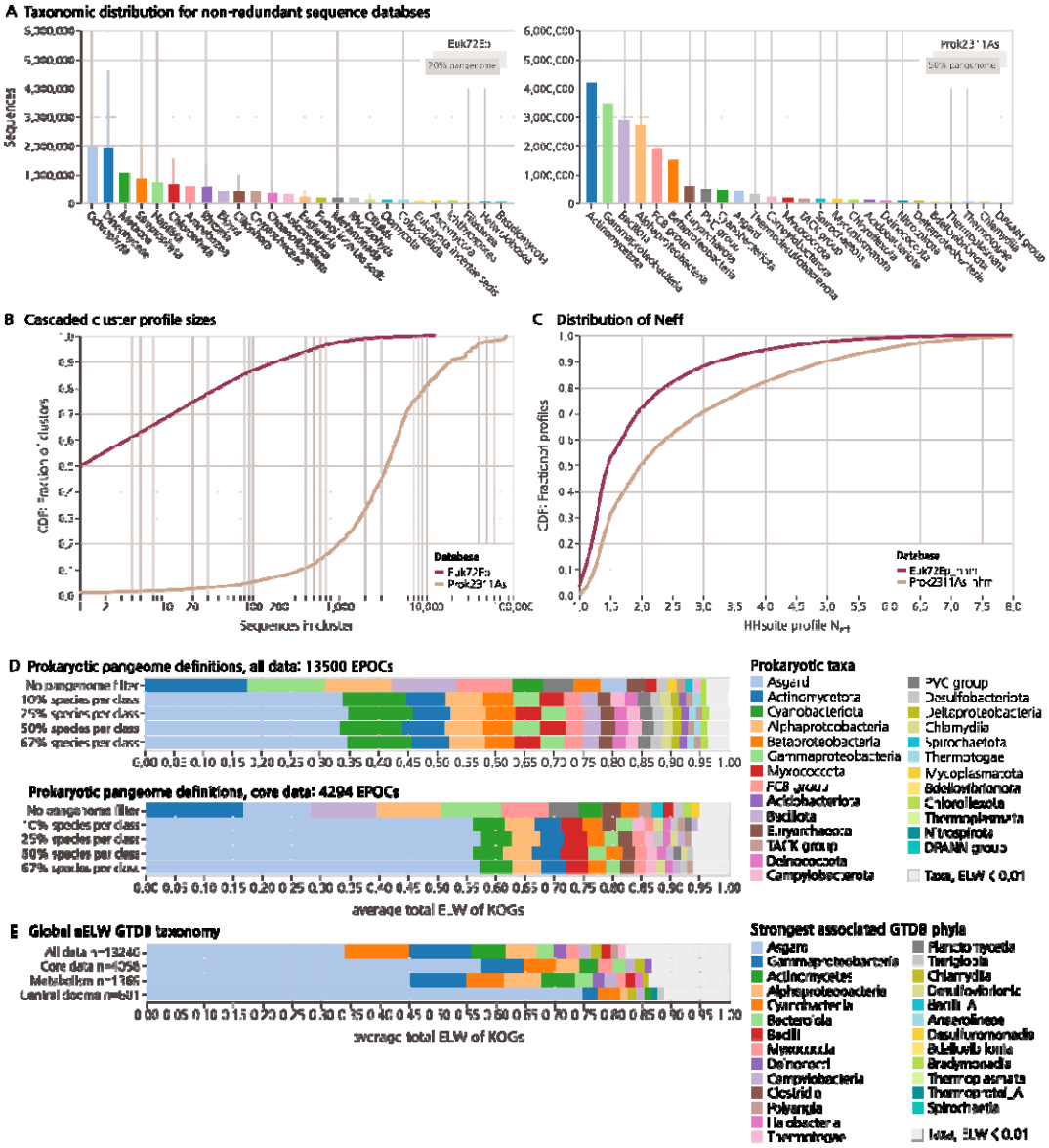
786 **References**

787 1 Embley, T. M. & Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440**, 623-630
788 (2006).
789 2 Vosseberg, J. *et al.* The emerging view on the origin and early evolution of eukaryotic cells.
790 *Nature* **633**, 295-305 (2024). https://doi.org:10.1038/s41586-024-07677-6
791 3 van der Giezen, M. Hydrogenosomes and mitosomes: conservation and evolution of functions. *J*
792 *Eukaryot Microbiol* **56**, 221-231 (2009). https://doi.org:JEU407 [pii] 10.1111/j.1550-
793 7408.2009.00407.x
794 4 Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes.
795 *Nature* **521**, 173-179 (2015). https://doi.org:10.1038/nature14447
796 5 Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular
797 complexity. *Nature* **541**, 353-358 (2017). https://doi.org:10.1038/nature21031
798 6 Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes.
799 *Nature* **593**, 553-557 (2021). https://doi.org:10.1038/s41586-021-03494-3
800 7 Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and the origin of
801 eukaryotes. *Nat Rev Microbiol* **15**, 711-723 (2017). https://doi.org:10.1038/nrmicro.2017.133
802 8 Eme, L. *et al.* Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes.
803 *Nature* **618**, 992-999 (2023). https://doi.org:10.1038/s41586-023-06186-2
804 9 Lopez-Garcia, P. & Moreira, D. The symbiotic origin of the eukaryotic cell. *C R Biol* **346**, 55-73
805 (2023). https://doi.org:10.5802/crbiol.118
806 10 Alberts, B. *et al. Molecular Biology of the Cell*. 7 edn, (W.W. Norton & Co, 2022).
807 11 Cavalier-Smith, T. Eukaryotes with no mitochondria. *Nature* **326**, 332-333 (1987).
808 https://doi.org:10.1038/326332a0
809 12 Sogin, M. L., Gunderson, J. H., Elwood, H. J., Alonso, R. A. & Peattie, D. A. Phylogenetic meaning
810 of the kingdom concept: an unusual ribosomal RNA from Giardia lamblia. *Science* **243**, 75-77
811 (1989). https://doi.org:10.1126/science.2911720
812 13 Sagan, L. On the origin of mitosing cells. *J Theor Biol* **14**, 255-274 (1967).
813 14 Martin, W., Hoffmeister, M., Rotte, C. & Henze, K. An overview of endosymbiotic models for the
814 origins of eukaryotes, their ATP-producing organelles (mitochondria and hydrogenosomes), and
815 their heterotrophic lifestyle. *Biol Chem* **382**, 1521-1539 (2001).
816 https://doi.org:10.1515/BC.2001.187
817 15 Lopez-Garcia, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited.
818 *Nat Microbiol* **5**, 655-667 (2020). https://doi.org:10.1038/s41564-020-0710-4
819 16 Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition. *Microbiol*
820 *Mol Biol Rev* **61**, 456-502 (1997). https://doi.org:10.1128/mmbr.61.4.456-502.1997
821 17 Rivera, M. C., Jain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct
822 gene classes. *Proc Natl Acad Sci U S A* **95**, 6239-6244 (1998).
823 https://doi.org:10.1073/pnas.95.11.6239
824 18 Ribeiro, S. & Golding, G. B. The mosaic nature of the eukaryotic nucleus. *Mol Biol Evol* **15**, 779-
825 788 (1998). https://doi.org:10.1093/oxfordjournals.molbev.a025983
826 19 Koonin, E. V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete
827 eukaryotic genomes. *Genome Biol* **5**, R7 (2004).
828 20 Imachi, H. *et al.* Promethearchaeum syntrophicum gen. nov., sp. nov., an anaerobic, obligately
829 syntrophic archaeon, the first isolate of the lineage 'Asgard' archaea, and proposal of the new
830 archaeal phylum Promethearchaeota phyl. nov. and kingdom Promethearchaeati regn. nov. *Int J*
831 *Syst Evol Microbiol* **74** (2024). https://doi.org:10.1099/ijsem.0.006435

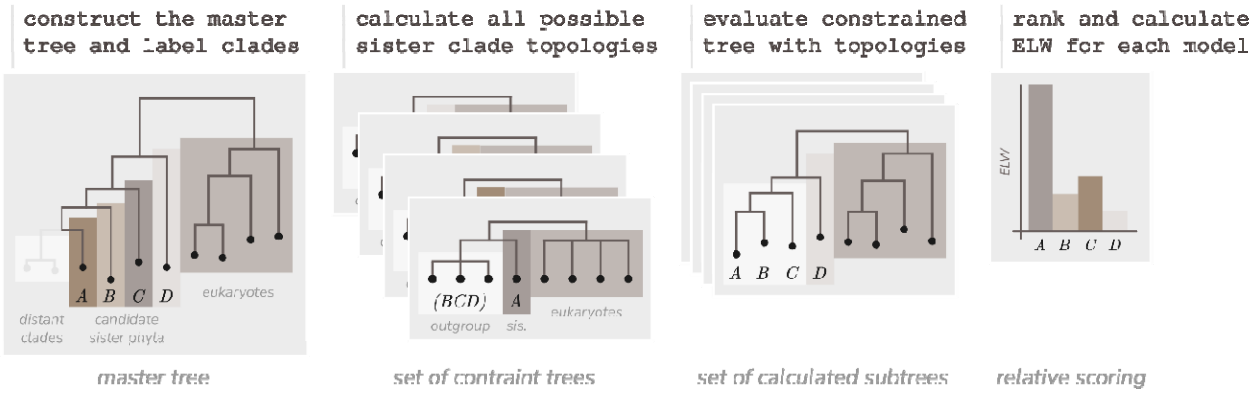
832 21 Lombard, J., Lopez-Garcia, P. & Moreira, D. The early evolution of lipid membranes and the
833 three domains of life. *Nat Rev Microbiol* **10**, 507-515 (2012).
834 https://doi.org:10.1038/nrmicro2815
835 22 Moreira, D. & Lopez-Garcia, P. Symbiosis between methanogenic archaea and delta-
836 proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J Mol Evol* **47**, 517-530
837 (1998). https://doi.org:10.1007/pl00006408
838 23 Krupovic, M., Dolja, V. V. & Koonin, E. V. The virome of the last eukaryotic common ancestor and
839 eukaryogenesis. *Nat Microbiol* **8**, 1008-1017 (2023). https://doi.org:10.1038/s41564-023-01378-
840 y
841 24 Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**,
842 519-525 (2020). https://doi.org:10.1038/s41586-019-1916-6
843 25 Lopez-Garcia, P. & Moreira, D. Cultured Asgard Archaea Shed Light on Eukaryogenesis. *Cell* **181**,
844 232-235 (2020). https://doi.org:10.1016/j.cell.2020.03.058
845 26 Pittis, A. A. & Gabaldon, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic
846 ancestry. *Nature* **531**, 101-104 (2016). https://doi.org:10.1038/nature16941
847 27 Gabaldon, T. Origin and Early Evolution of the Eukaryotic Cell. *Annu Rev Microbiol* **75**, 631-647
848 (2021). https://doi.org:10.1146/annurev-micro-090817-062213
849 28 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the
850 analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
851 https://doi.org:10.1038/nbt.3988
852 29 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein
853 annotation. *BMC Bioinformatics* **20**, 473 (2019). https://doi.org:10.1186/s12859-019-3019-7
854 30 Richter, D. J. *et al.* EukProt: A database of genome-scale predicted proteins across the diversity
855 of eukaryotes *Peer Community J* **2**, e56 (2022).
856 31 Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc*
857 *Biol Sci* **269**, 137-142 (2002). https://doi.org:10.1098/rspb.2001.1862
858 32 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on
859 genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017).
860 https://doi.org:10.1093/nar/gkw1092
861 33 Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of
862 sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022).
863 https://doi.org:10.1038/s41467-022-34630-w
864 34 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
865 the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020). https://doi.org:10.1093/molbev/msaa015
866 35 Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci U S A*
867 **112**, E693-699 (2015). https://doi.org:10.1073/pnas.1420657112
868 36 Al Jewari, C. & Baldauf, S. L. Conflict over the Eukaryote Root Resides in Strong Outliers, Mosaics
869 and Missing Data Sensitivity of Site-Specific (CAT) Mixture Models. *Syst Biol* **72**, 1-16 (2023).
870 https://doi.org:10.1093/sysbio/syac029
871 37 Paulick, M. G. & Bertozzi, C. R. The glycosylphosphatidylinositol anchor: a complex membrane-
872 anchoring structure for proteins. *Biochemistry* **47**, 6991-7000 (2008).
873 https://doi.org:10.1021/bi8006324
874 38 Stankeviciute, G. *et al.* Convergent evolution of bacterial ceramide synthesis. *Nat Chem Biol* **18**,
875 305-312 (2022). https://doi.org:10.1038/s41589-021-00948-7
876 39 Biran, A., Santos, T. C. B., Dingjan, T. & Futerman, A. H. The Sphinx and the egg: Evolutionary
877 enigmas of the (glyco)sphingolipid biosynthetic pathway. *Biochim Biophys Acta Mol Cell Biol*
878 *Lipids* **1869**, 159462 (2024). https://doi.org:10.1016/j.bbalip.2024.159462

879 40 Hoshino, Y. & Gaucher, E. A. On the Origin of Isoprenoid Biosynthesis. *Mol Biol Evol* **35**, 2185-
880 2197 (2018). https://doi.org:10.1093/molbev/msy120
881 41 Richards, T. A. & van der Giezen, M. Evolution of the Isd11-IscS complex reveals a single alpha-
882 proteobacterial endosymbiosis for all eukaryotes. *Mol Biol Evol* **23**, 1341-1344 (2006).
883 https://doi.org:10.1093/molbev/msl001
884 42 Cassier-Chauvat, C., Marceau, F., Farci, S., Ouchane, S. & Chauvat, F. The Glutathione System: A
885 Journey from Cyanobacteria to Higher Eukaryotes. *Antioxidants (Basel)* **12** (2023).
886 https://doi.org:10.3390/antiox12061199
887 43 Kassube, S. A. & Thoma, N. H. Structural insights into Fe-S protein biogenesis by the CIA
888 targeting complex. *Nat Struct Mol Biol* **27**, 735-742 (2020). https://doi.org:10.1038/s41594-020-
889 0454-0
890 44 Hedges, S. B. The origin and evolution of model organisms. *Nat Rev Genet* **3**, 838-849 (2002).
891 https://doi.org:10.1038/nrg929
892 45 Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications.
893 *Nat Ecol Evol* **5**, 92-100 (2021). https://doi.org:10.1038/s41559-020-01320-z
894 46 Caforio, A. *et al.* Converting Escherichia coli into an archaebacterium with a hybrid heterochiral
895 membrane. *Proc Natl Acad Sci U S A* **115**, 3704-3709 (2018).
896 https://doi.org:10.1073/pnas.1721604115
897 47 Hoekzema, M., Jiang, J. & Driessen, A. J. M. Optimizing Archaeal Lipid Biosynthesis in Escherichia
898 coli. *ACS Synth Biol* **13**, 2470-2479 (2024). https://doi.org:10.1021/acssynbio.4c00235
899 48 Tamarit, D. *et al.* A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal
900 viruses. *Nat Microbiol* **7**, 948-952 (2022). https://doi.org:10.1038/s41564-022-01122-y
901 49 Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard
902 archaeon. *Nature* **613**, 332-339 (2023). https://doi.org:10.1038/s41586-022-05550-y
903 50 Munoz-Gomez, S. A. *et al.* Site-and-branch-heterogeneous analyses of an expanded dataset
904 favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* **6**, 253-262 (2022).
905 https://doi.org:10.1038/s41559-021-01638-2
906 51 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
907 identification. *BMC Bioinformatics* **11**, 119 (2010). https://doi.org:10.1186/1471-2105-11-119
908 52 Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of
909 Phylogenomic Data. *Mol Biol Evol* **33**, 1635-1638 (2016).
910 https://doi.org:10.1093/molbev/msw046
911 53 Deorowicz, S., Debudaj-Grabysz, A. & Gudys, A. FAMSA: Fast and accurate multiple sequence
912 alignment of huge protein families. *Sci Rep* **6**, 33964 (2016). https://doi.org:10.1038/srep33964
913 54 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for
914 large alignments. *PLoS One* **5**, e9490 (2010). https://doi.org:10.1371/journal.pone.0009490
915 55 UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-
916 D489 (2021). https://doi.org:10.1093/nar/gkaa1100
917
918

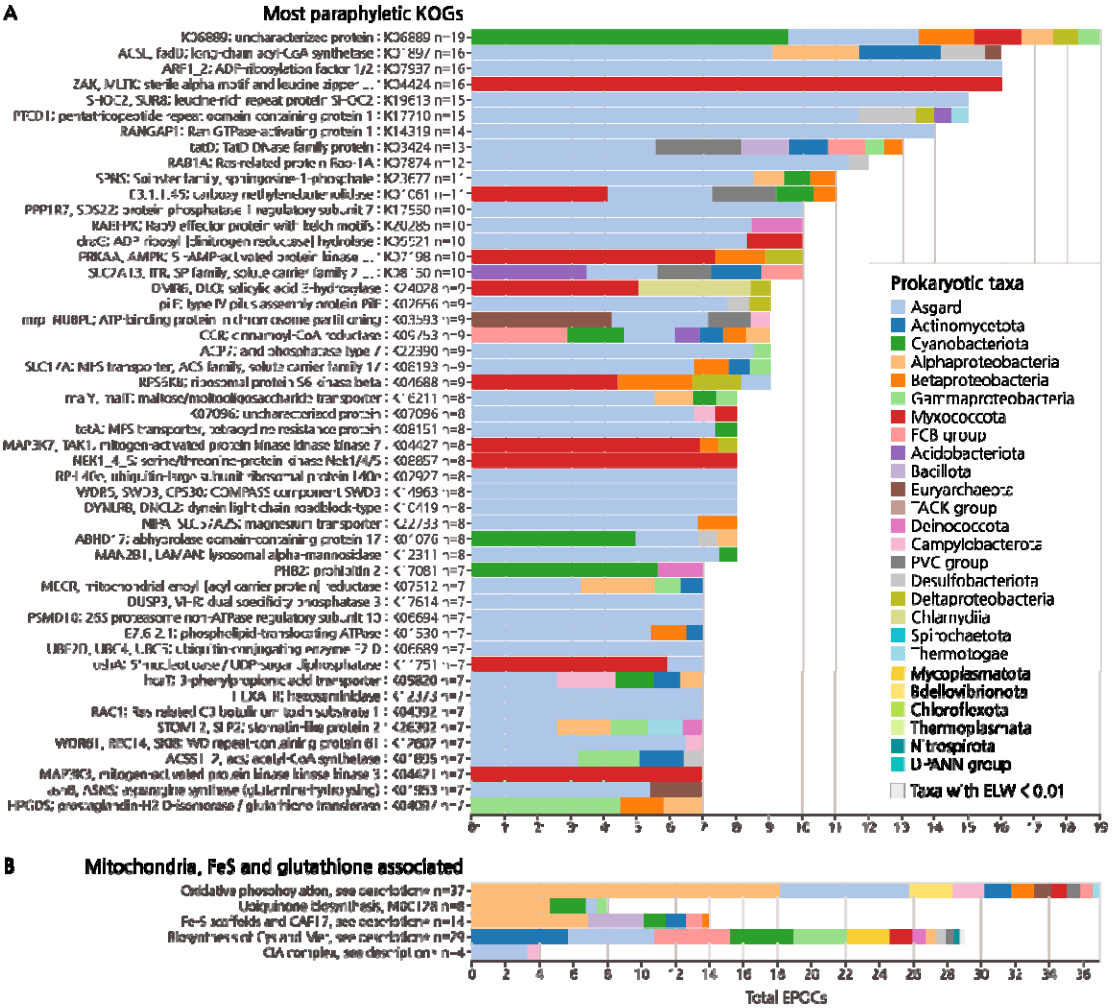
Extended Data



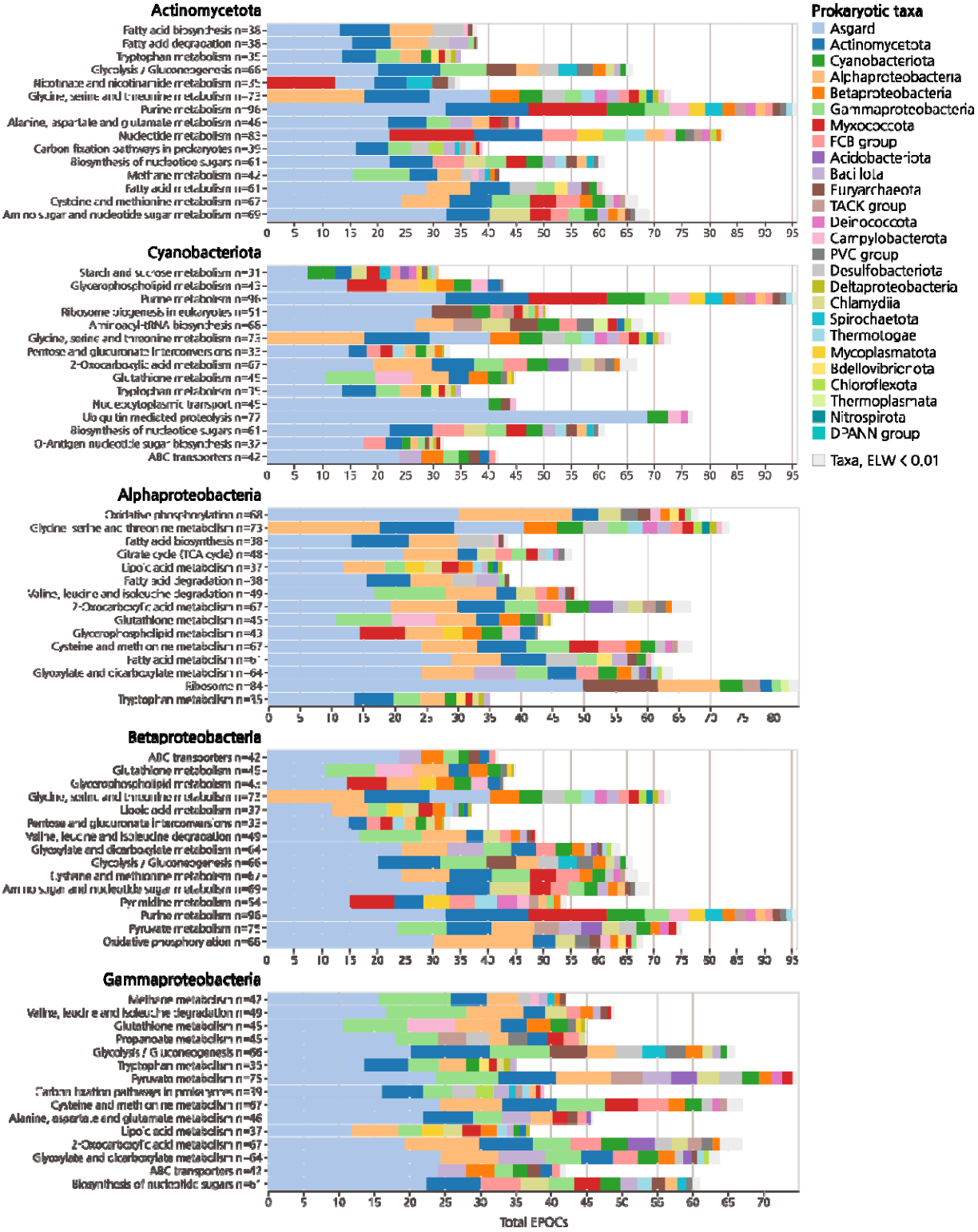
Supplementary Figure 1. Global database statistics. A) Distribution of taxonomic labels present within non-redundant Euk72Ep and Prok2111As (see Methods). Euk72Ep limited to only display taxa with more than 100.000 sequences (99.69% of sequences). **B)** CDF fraction of cluster size distribution following cascaded mmseqs profile/sequence clustering. **C)** CDF fraction of N_{eff} values as calculated by HH-suite for resulting profile databases. **D)** The effect of different pangenome regimes on final aELW values. Pangenomes are reconstructed enforcing a protein presence across 0, 25, 50, or 67% of species within clades and data shown for the full dataset of 14400 EPOCs or core dataset, as described in methods. 50% criteria marked in bold as it is used as reference for the core data present throughout the paper. **E)** Global aELW for data remapped to the GTDB taxonomy at the level of Phylum using marker genes (see Methods). Displaying the top 26 strongest individual contributing phyla out of 92 accounting for 96.4% of all ELW.



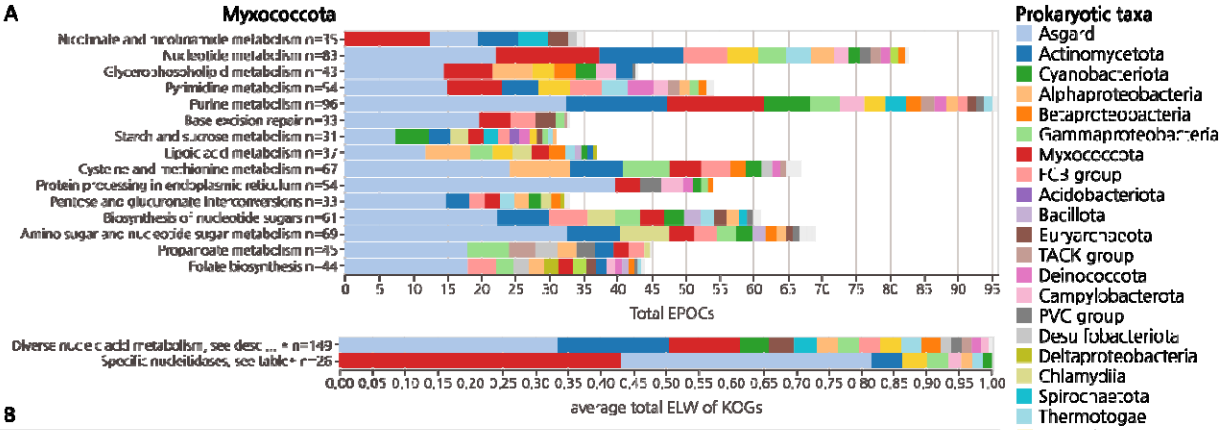
Supplementary Figure 2. Evolutionary hypothesis testing constraint trees and Expected Likelihood Weights. Overview of ELW calculation procedure. One unrooted master tree is constructed from all sequences within a EPOC and the closest n sister clades to the eukaryotic outgroup identified. A set of unrooted constraint trees are generated, each enforcing three clades, one eukaryotic, one sample sister and one outgroup with remaining prokaryotic sequences, is present. The set of subtrees are then constructed using IQtree, each forced to conform to a single guide tree. Log likelihood values for resulting trees are compared and evaluated using IQtree -z to produce confidence sets and Expected Likelihood Weight estimations.



Supplementary Figure 3. aELW breakdown of additional maps and pathways with prominent Asgard and Alphaproteobacterial contributions. A) Overview of most paralogous KOGs estimated by number of separate EPOCs traceable to the LECA. **B)** Oxidative phosphorylation (map00190) also contains entries for V-type ATPases which are excluded here as they are not part of mitochondrial metabolism. Fe-S scaffolds are not members of any map the manually curated set includes KOGs: K22063, K22072, K22068, K22073, K22070, K22071, K04043, K04044 and K04082. The CIA complex is likewise not grouped within BRITE and the manual set contains KOGs: K24730, K26403 and K15075.



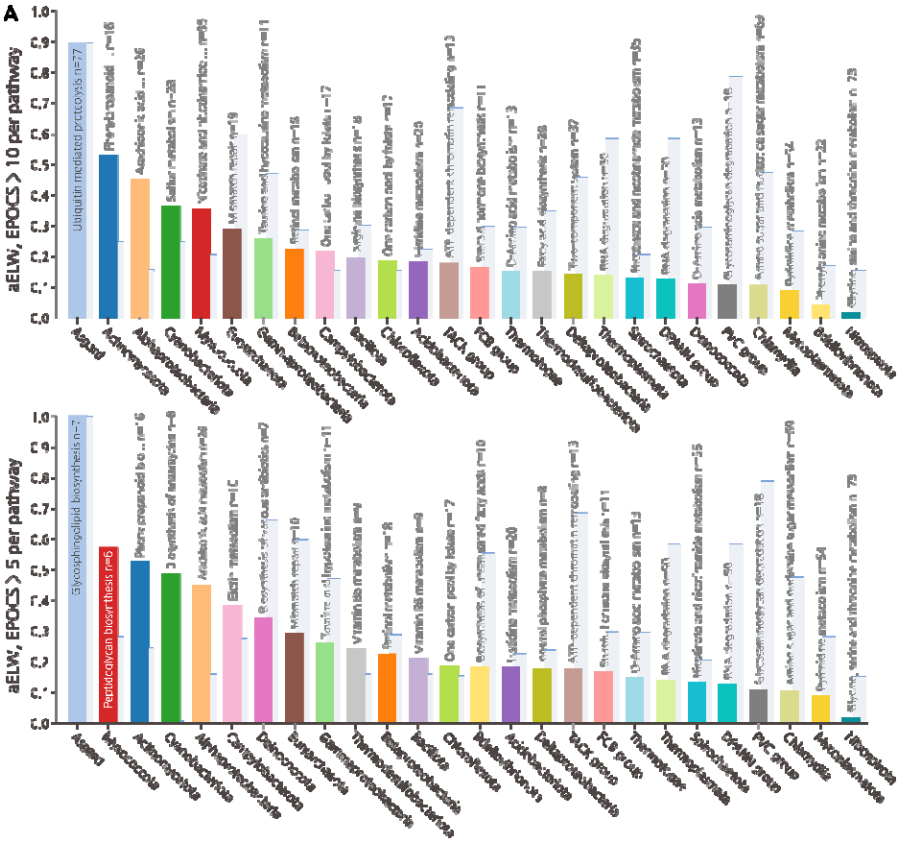
Supplementary Figure 4. Strongest per taxa association of diverse bacteria to eukaryotes. Top bacterial taxa associated taxa with aELW presented for their 15 most prominent pathways in KEGG Brite B. Pathways with more than 20 EPOCs shown.

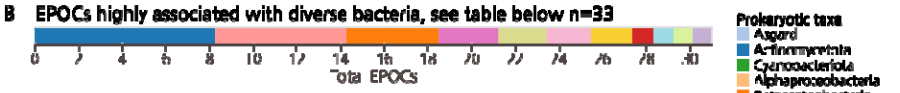


B

KOG	Gene Name	Description	EC *number(s)*
K10807	RRM1	ribonucleoside-diphosphate reductase subunit M1	1.17.4.1
K01517	ADPRM	manganese-dependent ADP-ribose/CDP-alcohol diphosphatase	3.6.1.13, 3.6.1.16, 3.6.1.53
K01081	E3.1.3.5	5'-nucleotidase	3.1.3.5
K01119	cpdB	2',3'-cyclic-nucleotide 2'-phosphodiesterase / 3'-nucleotidase	3.1.4.16, 3.1.3.6
K19970	NT5E, CD73	5'-nucleotidase	3.1.3.5
K03742	pncC	nicotinamide-nucleotide amidase	3.5.1.42
K03462	NAMPT	nicotinamide phosphoribosyltransferase	2.4.2.12
K11751	ushA	5'-nucleotidase / UDP-sugar diphosphatase	3.1.3.5, 3.6.1.45
K01525	apaH	bis(5' nucleosyl) tetraphosphatase (symmetrical)	3.6.1.41
K03651	cpdA	3',5'-cyclic-AMP phosphodiesterase	3.1.4.53
K12410	**cobB, sirN, npdA**	**NAD-dependent protein deacetylase/lipoamidase**	**2.3.1.286, 2.3.1.313**

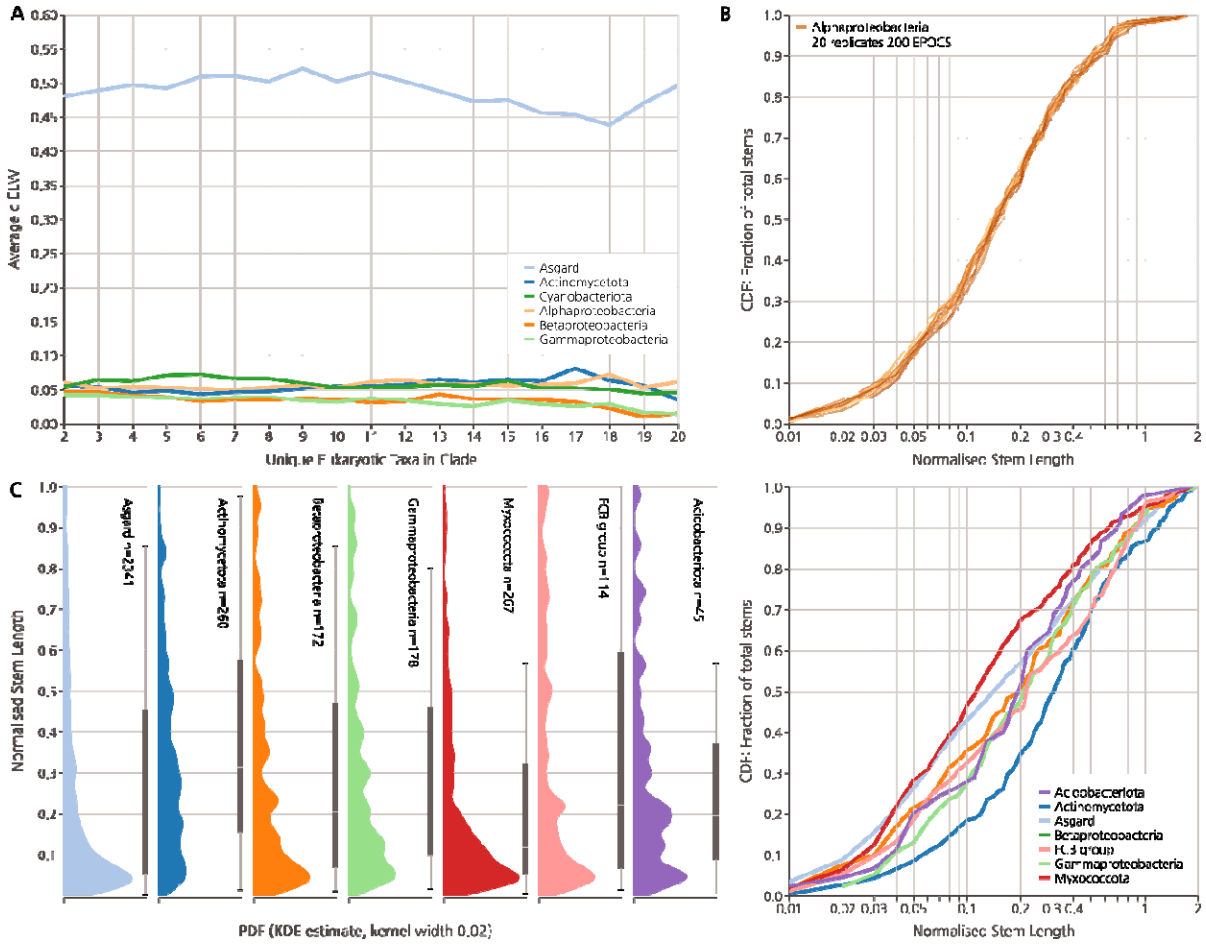
Supplementary Figure 5. Sparse Myxococcota association with eukaryotes. A) Top, aELW breakdown of enriched Myxococcota pathways showing primary association with nucleotide synthesis. Bottom, curated set of Myxococcota enriched EPOCs containing general nucleic acid metabolism and modification pathways corresponding to KEGG map00760, map01232, map00240 and map00230 with notable presence of Myxococcota. Further selection of only those KOGs within the above pathways which include Myxococcota. **B)** List of KOGs with Myxococcota presence within the curated set of nucleotide synthesis highlighting phosphatases and phosphotransferases.



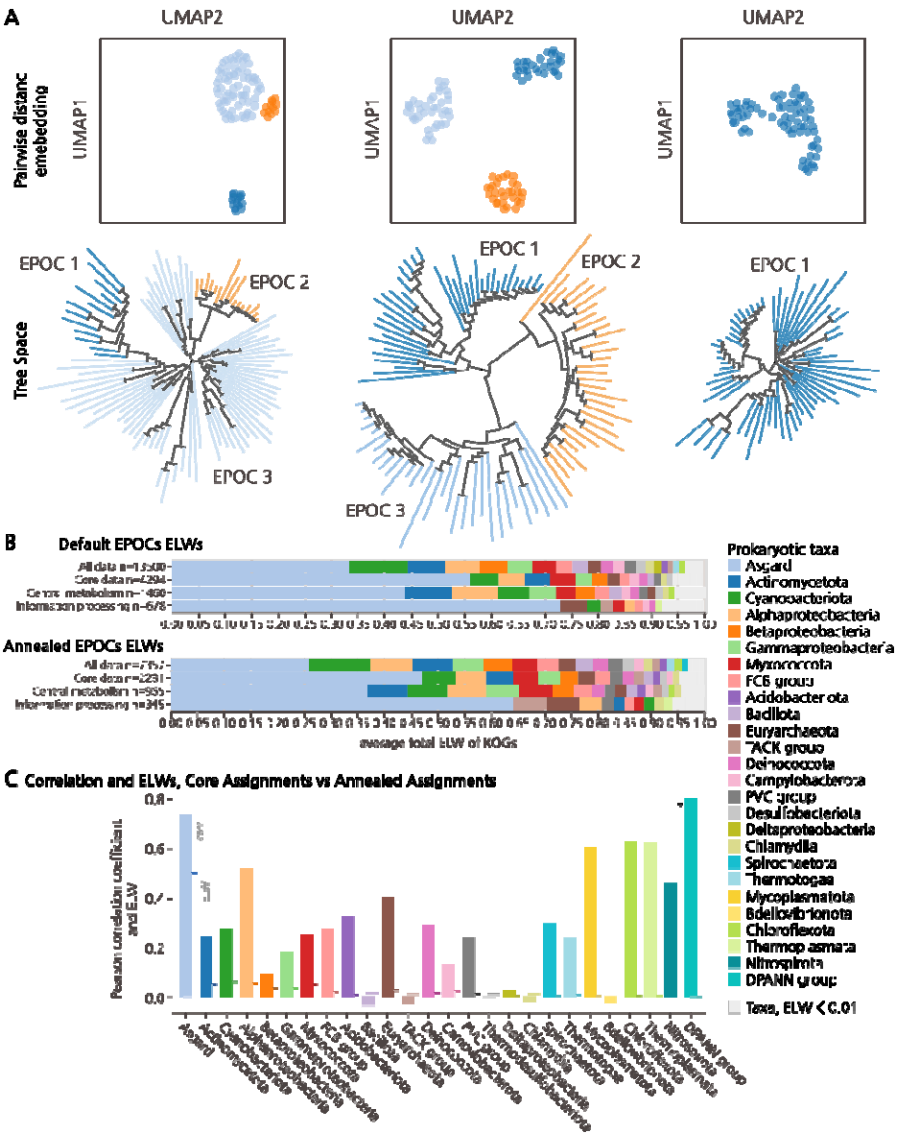


KOG	Gene Name	Description	EC number(s)
K02564	nagB, GNPDA	glucosamine-6-phosphate deaminase	3.5.99.6
K12862	PLRG1, PRL1, PRP46	pleiotropic regulator 1	
K07198	PRKAA, AMPK	5'-AMP-activated protein kinase, catalytic alpha subunit	2.7.11.31
K10830	ERCC8, CKN1, CSA	DNA excision repair protein ERCC-8	
K06972	PITRM1, PrcP, CYM1	presequence protease	3.4.24.-
K05967	K05967	uncharacterized protein	
K03781	catE, CAT, cat3, srpA	catalase	1.11.1.6
K04568	RPS6KB	ribosomal protein S6 kinase beta	2.7.11.1
K01908	ACSS3, prpE	propionyl-CoA synthetase	6.2.1.17
K12187	STRAP, UNRIP	serine-threonine kinase receptor-associated protein	
K06013	STE24	STE24 endopeptidase	3.4.24.84
K14085	ALDH7A1	aldehyde dehydrogenase family 7 member A1	1.2.1.31, 1.2.1.8, 1.2.1.3
K00547	MTR, MS	methionine synthase reductase	1.16.1.8
K00023	phbB	acetoacetyl-CoA reductase	1.1.1.36
K14450	PHO87_91	phosphate transporter	
K08956	AFG3	AFG3 family protein	3.4.24.-
K00261	GLUD1_2, gdhA	glutamate dehydrogenase (NAD(P)+)	1.4.1.3
K12450	RHM	UDP-glucose 4,6-dehydratase	4.2.1.76
K00281	GLDC, gcvP	glycine cleavage system P protein (glycine dehydrogenase)	1.4.4.2
K11141	ENPEP, CD249	glutamyl aminopeptidase	3.4.11.7

Supplementary Figure 6. KOGs of strong bacterial association. A) Same as main figure 4 under less stringent selection criteria indicating most prominent prokaryotic pathways with a coverage of at least 10 and 5 EPOCs, aELW shown relative to that of Asgard for the same pathway in pale blue. **B)** Individual EPOCs strongly associated with diverse prokaryotic taxa under stringent selection criteria. EPOCs must contain a eukaryotic outgroup with at least 15 unique taxa encompassing the LECA, have a prokaryotic sister clade with an ELW of > 0.7 which is not derived from Asgard, Euryarchaeota, the TACK group, DPANN group, Cyanobacteriota or Alphaproteobacteria and contain more than 20 sequences. The EPOCs which meet these criteria are highlighted with respective KOGs indicated, aELW displayed.



Supplementary Figure 7: **Clade scope definition and stem length distribution for diverse bacteria compared to Asgard**. **A)** Global aELW dependence on choice of clade size cutoff using core data. Estimated aELW are independent on choice of eukaryotic clade scope. **B)** 20 full technical replicates, including muscle5 alignment, IQtree2 calculations and constraint tree generation, of 175 randomly sampled Alphaproteobacterial stems showing reproducibility of stem length distributions. Compared with distributions drawn from Asgard and Cyanobacterial stems as per Figure 5. **C)** Left, Probability Density Function (PDF) of normalized stem lengths taken from samples of individual core eukaryotic genes associated with prokaryotic taxa, obtained through KDE. Thin bar indicates 5th percentile ranges, thick bar includes 25th percentile ranges with median indicated. Right, fractional cumulative distribution functions of normalized stem lengths.



Supplementary Figure 8: **Annealed cluster EPOCS A)** Examples of annealed clades defined based on embedded and clustered pairwise distance matrices of phylogenetic trees. Top, embedding of pairwise leaf distance of phylogenetic tree, bottom. Clustered by HDBSCAN allowing single cluster partitions as shown, left. Partitions form monophyletic subsets of well partitioned branched. **B)** Resulting top level aELW values derived from the core data clusters as well as annealed EPOCs showing broad correlation between different core and annealed partitioning schemes. **C)** Correlation of aELW vectors calculated for all Prokaryotic taxa across all KEGG maps using the annealed and core EPOCs indicating good agreement for Asgard assignments but greater variance between prokaryotic taxa. Overlaid on global aELW values. Taxa for which aELW is less than 0.01 are found to be associated only for very few pathways and as such correlate strongly, notably DPANN has an aELW of 0 and do not contribute to any pathways, and therefore with a correlation of 1.0 between the two reconstructions.

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 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
4 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
5 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
6 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
7 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
8 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
9 Alberts, B. *et al. Molecular Biology of the Cell*. 7 edn, (W.W. Norton & Co, 2022).
10 Cavalier-Smith, T. Eukaryotes with no mitochondria. *Nature* **326**, 332-333 (1987). https://doi.org:10.1038/326332a0
11 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
12 van der Giezen, M. Hydrogenosomes and mitosomes: conservation and evolution of functions. *J Eukaryot Microbiol* **56**, 221-231 (2009). https://doi.org:JEU407 [pii] 10.1111/j.1550-7408.2009.00407.x
13 Sagan, L. On the origin of mitosing cells. *J Theor Biol* **14**, 255-274 (1967).
14 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
15 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
16 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
17 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
18 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
19 Koonin, E. V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* **5**, R7 (2004).
20 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
21 Tamarit, D. *et al.* Description of Asgardarchaeum abyssi gen. nov. spec. nov., a novel species within the class Asgardarchaeia and phylum Asgardarchaeota in accordance with the SeqCode. *Syst Appl Microbiol* **47**, 126525 (2024). https://doi.org:10.1016/j.syapm.2024.126525

1046 22 Wu, F. *et al.* Unique mobile elements and scalable gene flow at the prokaryote-eukaryote
1047 boundary revealed by circularized Asgard archaea genomes. *Nat Microbiol* **7**, 200-212 (2022).
1048 https://doi.org:10.1038/s41564-021-01039-y
1049 23 Valentin-Alvarado, L. E. *et al.* Complete genomes of Asgard archaea reveal diverse integrated
1050 and mobile genetic elements. *Genome Res* **34**, 1595-1609 (2024).
1051 https://doi.org:10.1101/gr.279480.124
1052 24 Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard
1053 archaeon. *Nature* **613**, 332-339 (2023). https://doi.org:10.1038/s41586-022-05550-y
1054 25 Tran, L. T., Akil, C., Senju, Y. & Robinson, R. C. The eukaryotic-like characteristics of small GTPase,
1055 roadblock and TRAPPC3 proteins from Asgard archaea. *Commun Biol* **7**, 273 (2024).
1056 https://doi.org:10.1038/s42003-024-05888-1
1057 26 Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications.
1058 *Nat Ecol Evol* **5**, 92-100 (2021). https://doi.org:10.1038/s41559-020-01320-z
1059 27 Richards, T. A. *et al.* Reconstructing the last common ancestor of all eukaryotes. *PLoS Biol* **22**,
1060 e3002917 (2024). https://doi.org:10.1371/journal.pbio.3002917
1061 28 Lombard, J., Lopez-Garcia, P. & Moreira, D. The early evolution of lipid membranes and the
1062 three domains of life. *Nat Rev Microbiol* **10**, 507-515 (2012).
1063 https://doi.org:10.1038/nrmicro2815
1064 29 Moreira, D. & Lopez-Garcia, P. Symbiosis between methanogenic archaea and delta-
1065 proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J Mol Evol* **47**, 517-530
1066 (1998). https://doi.org:10.1007/pl00006408
1067 30 Krupovic, M., Dolja, V. V. & Koonin, E. V. The virome of the last eukaryotic common ancestor and
1068 eukaryogenesis. *Nat Microbiol* **8**, 1008-1017 (2023). https://doi.org:10.1038/s41564-023-01378-
1069 y
1070 31 Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**,
1071 519-525 (2020). https://doi.org:10.1038/s41586-019-1916-6
1072 32 Lopez-Garcia, P. & Moreira, D. Cultured Asgard Archaea Shed Light on Eukaryogenesis. *Cell* **181**,
1073 232-235 (2020). https://doi.org:10.1016/j.cell.2020.03.058
1074 33 Gabaldon, T. Relative timing of mitochondrial endosymbiosis and the "pre-mitochondrial
1075 symbioses" hypothesis. *IUBMB Life* **70**, 1188-1196 (2018). https://doi.org:10.1002/iub.1950
1076 34 Gabaldon, T. Origin and Early Evolution of the Eukaryotic Cell. *Annu Rev Microbiol* **75**, 631-647
1077 (2021). https://doi.org:10.1146/annurev-micro-090817-062213
1078 35 Pittis, A. A. & Gabaldon, T. Late acquisition of mitochondria by a host with chimaeric prokaryotic
1079 ancestry. *Nature* **531**, 101-104 (2016). https://doi.org:10.1038/nature16941
1080 36 Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat*
1081 *Rev Genet* **16**, 472-482 (2015). https://doi.org:10.1038/nrg3962
1082 37 Artamonova, II, Lappi, T., Zudina, L. & Mushegian, A. R. Prokaryotic genes in eukaryotic genome
1083 sequences: when to infer horizontal gene transfer and when to suspect an actual microbe.
1084 *Environ Microbiol* **17**, 2203-2208 (2015). https://doi.org:10.1111/1462-2920.12854
1085 38 Keeling, P. J. Horizontal gene transfer in eukaryotes: aligning theory with data. *Nat Rev Genet*
1086 **25**, 416-430 (2024). https://doi.org:10.1038/s41576-023-00688-5
1087 39 Wittmers, F. *et al.* Symbionts of predatory protists are widespread in the oceans and related to
1088 animal pathogens. *Cell Host Microbe* **33**, 182-199 e187 (2025).
1089 https://doi.org:10.1016/j.chom.2025.01.009
1090 40 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein
1091 annotation. *BMC Bioinformatics* **20**, 473 (2019). https://doi.org:10.1186/s12859-019-3019-7
1092 41 Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc*
1093 *Biol Sci* **269**, 137-142 (2002). https://doi.org:10.1098/rspb.2001.1862

1094 42 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on
1095 genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353-D361 (2017).
1096 https://doi.org:10.1093/nar/gkw1092
1097 43 Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of
1098 sequence homology and phylogeny. *Nat Commun* **13**, 6968 (2022).
1099 https://doi.org:10.1038/s41467-022-34630-w
1100 44 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in
1101 the Genomic Era. *Mol Biol Evol* **37**, 1530-1534 (2020). https://doi.org:10.1093/molbev/msaa015
1102 45 Al Jewari, C. & Baldauf, S. L. Conflict over the Eukaryote Root Resides in Strong Outliers, Mosaics
1103 and Missing Data Sensitivity of Site-Specific (CAT) Mixture Models. *Syst Biol* **72**, 1-16 (2023).
1104 https://doi.org:10.1093/sysbio/syac029
1105 46 Derelle, R. *et al.* Bacterial proteins pinpoint a single eukaryotic root. *Proc Natl Acad Sci U S A*
1106 **112**, E693-699 (2015). https://doi.org:10.1073/pnas.1420657112
1107 47 Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a
1108 phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic*
1109 *Acids Res* **50**, D785-D794 (2022). https://doi.org:10.1093/nar/gkab776
1110 48 Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G. & Koonin, E. V. Ancestral paralogs and
1111 pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res* **33**,
1112 4626-4638 (2005). https://doi.org:33/14/4626 [pii]
1113 10.1093/nar/gki775
1114 49 Keeling, P. J. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution.
1115 *Annu Rev Plant Biol* **64**, 583-607 (2013). https://doi.org:10.1146/annurev-arplant-050312-
1116 120144
1117 50 Archibald, J. M. Genomic perspectives on the birth and spread of plastids. *Proc Natl Acad Sci U S*
1118 *A* **112**, 10147-10153 (2015). https://doi.org:10.1073/pnas.1421374112
1119 51 Ponce-Toledo, R. I., Lopez-Garcia, P. & Moreira, D. Horizontal and endosymbiotic gene transfer
1120 in early plastid evolution. *New Phytol* **224**, 618-624 (2019). https://doi.org:10.1111/nph.15965
1121 52 Paulick, M. G. & Bertozzi, C. R. The glycosylphosphatidylinositol anchor: a complex membrane-
1122 anchoring structure for proteins. *Biochemistry* **47**, 6991-7000 (2008).
1123 https://doi.org:10.1021/bi8006324
1124 53 Biran, A., Santos, T. C. B., Dingjan, T. & Futerman, A. H. The Sphinx and the egg: Evolutionary
1125 enigmas of the (glyco)sphingolipid biosynthetic pathway. *Biochim Biophys Acta Mol Cell Biol*
1126 *Lipids* **1869**, 159462 (2024). https://doi.org:10.1016/j.bbalip.2024.159462
1127 54 Stankeviciute, G. *et al.* Convergent evolution of bacterial ceramide synthesis. *Nat Chem Biol* **18**,
1128 305-312 (2022). https://doi.org:10.1038/s41589-021-00948-7
1129 55 Hoshino, Y. & Gaucher, E. A. On the Origin of Isoprenoid Biosynthesis. *Mol Biol Evol* **35**, 2185-
1130 2197 (2018). https://doi.org:10.1093/molbev/msy120
1131 56 Santana-Molina, C., Williams, T. A., Snel, B. & Spang, A. Chimeric origins and dynamic evolution
1132 of central carbon metabolism in eukaryotes. *Nat Ecol Evol* (2025).
1133 https://doi.org:10.1038/s41559-025-02648-0
1134 57 Richards, T. A. & van der Giezen, M. Evolution of the Isd11-IscS complex reveals a single alpha-
1135 proteobacterial endosymbiosis for all eukaryotes. *Mol Biol Evol* **23**, 1341-1344 (2006).
1136 https://doi.org:10.1093/molbev/msl001
1137 58 Cassier-Chauvat, C., Marceau, F., Farci, S., Ouchane, S. & Chauvat, F. The Glutathione System: A
1138 Journey from Cyanobacteria to Higher Eukaryotes. *Antioxidants (Basel)* **12** (2023).
1139 https://doi.org:10.3390/antiox12061199

1140 59 Kassube, S. A. & Thoma, N. H. Structural insights into Fe-S protein biogenesis by the CIA
1141 targeting complex. *Nat Struct Mol Biol* **27**, 735-742 (2020). https://doi.org:10.1038/s41594-020-
1142 0454-0
1143 60 Hoshino, Y. & Gaucher, E. A. Evolution of bacterial steroid biosynthesis and its impact on
1144 eukaryogenesis. *Proc Natl Acad Sci U S A* **118** (2021). https://doi.org:10.1073/pnas.2101276118
1145 61 Donoghue, P. C. J. *et al.* Defining eukaryotes to dissect eukaryogenesis. *Curr Biol* **33**, R919-R929
1146 (2023). https://doi.org:10.1016/j.cub.2023.07.048
1147 62 Hedges, S. B. The origin and evolution of model organisms. *Nat Rev Genet* **3**, 838-849 (2002).
1148 https://doi.org:10.1038/nrg929
1149 63 Koonin, E. V. & Yutin, N. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA
1150 viruses. *Intervirology* **53**, 284-292 (2010). https://doi.org:000312913 [pii]
1151 10.1159/000312913
1152 64 Caforio, A. *et al.* Converting Escherichia coli into an archaebacterium with a hybrid heterochiral
1153 membrane. *Proc Natl Acad Sci U S A* **115**, 3704-3709 (2018).
1154 https://doi.org:10.1073/pnas.1721604115
1155 65 Hoekzema, M., Jiang, J. & Driessen, A. J. M. Optimizing Archaeal Lipid Biosynthesis in Escherichia
1156 coli. *ACS Synth Biol* **13**, 2470-2479 (2024). https://doi.org:10.1021/acssynbio.4c00235
1157 66 Martin, W. & Koonin, E. V. Introns and the origin of nucleus-cytosol compartmentalization.
1158 *Nature* **440**, 41-45 (2006). https://doi.org:nature04531 [pii]
1159 10.1038/nature04531
1160 67 Lopez-Garcia, P. & Moreira, D. Selective forces for the origin of the eukaryotic nucleus. *Bioessays*
1161 **28**, 525-533 (2006). https://doi.org:10.1002/bies.20413
1162 68 Padilla-Mejia, N. E., Makarov, A. A., Barlow, L. D., Butterfield, E. R. & Field, M. C. Evolution and
1163 diversification of the nuclear envelope. *Nucleus* **12**, 21-41 (2021).
1164 https://doi.org:10.1080/19491034.2021.1874135
1165 69 Mans, B. J., Anantharaman, V., Aravind, L. & Koonin, E. V. Comparative genomics, evolution and
1166 origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* **3**, 1612-1637 (2004).
1167 https://doi.org:1316 [pii]
1168 70 Tamarit, D. *et al.* A closed Candidatus Odinarchaeum chromosome exposes Asgard archaeal
1169 viruses. *Nat Microbiol* **7**, 948-952 (2022). https://doi.org:10.1038/s41564-022-01122-y
1170 71 Munoz-Gomez, S. A. *et al.* Site-and-branch-heterogeneous analyses of an expanded dataset
1171 favour mitochondria as sister to known Alphaproteobacteria. *Nat Ecol Evol* **6**, 253-262 (2022).
1172 https://doi.org:10.1038/s41559-021-01638-2
1173 72 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site
1174 identification. *BMC Bioinformatics* **11**, 119 (2010). https://doi.org:10.1186/1471-2105-11-119
1175 73 Richter, D. J. *et al.* EukProt: A database of genome-scale predicted proteins across the diversity
1176 of eukaryotes *Peer Community J* **2**, e56 (2022).
1177 74 Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of
1178 Phylogenomic Data. *Mol Biol Evol* **33**, 1635-1638 (2016).
1179 https://doi.org:10.1093/molbev/msw046
1180 75 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the
1181 analysis of massive data sets. *Nat Biotechnol* **35**, 1026-1028 (2017).
1182 https://doi.org:10.1038/nbt.3988
1183 76 Deorowicz, S., Debudaj-Grabysz, A. & Gudys, A. FAMSA: Fast and accurate multiple sequence
1184 alignment of huge protein families. *Sci Rep* **6**, 33964 (2016). https://doi.org:10.1038/srep33964
1185 77 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for
1186 large alignments. *PLoS One* **5**, e9490 (2010). https://doi.org:10.1371/journal.pone.0009490

1187 78 UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480-
1188 D489 (2021). https://doi.org:10.1093/nar/gkaa1100
1189 79 Chen, H., Zheng, F., Feng, X., et al. Engineering archaeal membrane-spanning lipid GDGT
1190 biosynthesis in bacteria: Implications for early life membrane transformations. mLife. **4**:193-204
1191 (2025) https://doi.org:10.1002/mlf2.70001
1192