Phylogenomic analyses reveal that Panguiarchaeum is a clade of genome-reduced Asgard archaea

Wen-Cong Huang[1,2], Maraike Probst[3], Zheng-Shuang Hua[4], Lénárd L. Szánthó[5], Gergely J. Szöllősi[5,6], Thijs J. G. Ettema[7], Christian Rinke[3], Tom A. Williams[8*], Anja Spang[1,2*]

1. Department of Marine Microbiology and Biogeochemistry, NIOZ, Royal Netherlands Institute for Sea Research, 1790 AB Den Burg, The Netherlands
2. Department of Evolutionary & Population Biology, Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, 1090 GE Amsterdam, The Netherlands
3. Faculty of Biology, Department of Microbiology, University of Innsbruck, 6020 Innsbruck, Austria
4. Chinese Academy of Sciences Key Laboratory of Urban Pollutant Conversion, Department of Environmental Science and Engineering, University of Science and Technology of China, 230026 Hefei, China
5. Model-Based Evolutionary Genomics Unit, Okinawa Institute of Science and Technology Graduate University, 904-0495 Okinawa, Japan
6. Institute of Evolution, HUN-REN Center for Ecological Research, H-1121 Budapest, Hungary
7. Laboratory of Microbiology, Wageningen University & Research, Wageningen, The Netherlands
8. Bristol Palaeobiology Group, School of Biological Sciences, University of Bristol, BS8 1TQ, Bristol, UK

*Corresponding authors: Anja Spang (anja.spang@nioz.nl) and Tom A. Williams (tom.a.williams@bristol.ac.uk)

Abstract

The Asgard archaea are a diverse archaeal phylum that includes the host lineage from which eukaryotes evolved. Due to the importance of the Asgard archaea for our understanding of cellular evolution and eukaryotic complexity, recent efforts have focused on better characterizing the genomic diversity of this lineage and its relatives from the TACK archaea. Newly discovered lineages, the Njordarchaeales and Panguiarchaeales, are difficult to place, branching either with Korarchaeota or Asgard archaea. The phylogenetic position of these taxa is important because they include genome-reduced and potentially symbiotic lineages, which may help to inform our understanding of genome evolution and the evolution of symbiotic interactions within Asgard archaea and their relatives. To resolve the placement of Njordarchaeales and Panguiarchaeales in the archaeal tree, we performed a range of phylogenetic analyses revealing that the Njordarchaeales and Pangiuarchaeales constitute the newly-identified class Njordarchaeia within the Asgard archaea. Members of the Njordarchaeia exhibit hallmarks of adaptations to (hyper-) thermophilic lifestyles, which may contribute conflicting signals to phylogenetic analyses. *Panguiarchaeum* appears to be metabolically distinct from its relatives, displaying reduced metabolic potential and various auxotrophies. Gene tree-species tree reconciliation recovers a complex common ancestor of Asgard archaea that encoded the Wood Ljungdahl pathway, however the subsequent loss of this pathway during the reductive evolution of *Panguiarchaeum* may have been associated with the switch to a symbiotic lifestyle based on H_2 syntrophy. Our analyses identify *Panguiarchaeum* as the first described Asgard archaea lineage with streamlined genomes that show indications for a symbiotic lifestyle.

Introduction

Asgard archaea are a diverse phylum from which the archaeal ancestor of eukaryotes emerged[1–4]. The genomes of Asgard archaea[1–6] encode eukaryotic signature proteins (ESPs)[7], which are homologs of eukaryotic proteins that underlie cellular complexity. Experimental evidence suggests that at least some ESPs of Asgard archaea have functions similar to those in eukaryotes[8–12]. Genomic predictions and the cultivation of the first representatives of the Asgard archaea suggest that at least some members form symbiotic, (that is, syntrophic), interactions with other cells[13,14]. Indeed, eukaryotes are thought to have evolved from one such interaction[7,15–23]. The phylogeny and genomic features of Asgard archaea and related groups therefore provide an important framework for understanding the origin of eukaryotic cells and the evolution of prokaryotic symbioses[1–4,24]. A major current challenge lies in resolving the evolutionary relationships among Asgard archaea and their relatives, including Thermoproteota (formerly Thaumarchaeota, Aigarchaeota, and Crenarchaeota) and Korarchaeota[25,26]. Early analyses supported a clade of “TACK” Archaea[27] (initially comprising the phyla Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Koarchaeota now classified as class-level lineages within the Thermoproteota) as the sister lineage to the Asgard archaea[1,2,28]. However, more recent phylogenetic analyses taking advantage of improved genomic sampling have instead placed the Korarchaeota at the base of a monophyletic Thermoproteota and Asgard clade[26,29].

Two additional lineages have recently been described that appear to be part of a broader Asgard/Thermoproteota/Korarchaeota clade, although their phylogenetic affinities remain uncertain: the Njordarchaeales[4,30,31] and the Panguiarchaeales[32]. Njordarchaeales have been recovered either as sister to Korarchaeota[31] - with whom they share a thermophilic lifestyle - or, intriguingly, within Asgard archaea, either as sister to the eukaryotic nuclear lineage[30], or an order within the Heimdallarchaeia[4], depending on the dataset and phylogenetic method used. Similarly, the recently-reported Panguiarchaeales were reported to branch with Korarchaeota[32]. Panguiarchaeales were proposed to be symbiotic[32] because they lack certain genes for archaeal lipid, purine and amino acid biosynthesis pathways. However, the authors could not identify direct evidence for reductive genome evolution in the investigated

metagenome-assembled genomes (MAGs), which did not differ significantly from other Korarchaeota with which they were affiliated[32]. Reduced metabolic capabilities are hallmarks of the DPANN archaea, which represent diverse phylum-level lineages comprised of putative symbionts[29,33,34], but are otherwise uncommon among archaea. Thus, resolving the evolutionary relationships within the Asgard/Thermoproteota/Korarchaeota clade, and the positions of Njordarchaeales and Panguiarchaeales within them, is key to understanding the evolutionary trajectories of genome reduction and symbiotic associations in Archaea as well as for placing the origin of eukaryotic cells in the broader context of archaeal evolution.

Here, we applied a variety of comparative genomics, phylogenomic, and network approaches to determine the phylogenetic placement of Njordarchaeales and Panguiarchaeales within the archaeal tree of life and to assess the genomic potential and proposed symbiotic lifestyle of the latter. Our analyses reveal that the previously described Njordarchaeales and Panguiarchaeales orders are order- or family-level lineages which together form the new class Njordarchaeia within the Asgard archaea rather than the Korarchaeota. While all Njordarchaeia encode a similar set of ESPs, *Panguiarchaeaceae* MAGs show indications of reductive genome evolution and auxotrophies, suggesting that they may be dependent on partner organisms or other community members to supplement their growth. Indeed, our co-occurrence analyses reveal a higher-than-expected frequency of association between *Panguiarchaeum* and members of the Thermoproteota. Overall, this study highlights the need to implement careful phylogenetic approaches for accurately placing archaeal lineages comprising symbionts or extremophiles in a resolved phylogeny. It further provides the first example of a putative symbiotic lineage within the Asgard archaea that has experienced gene loss rather than genome expansion.

Results

***Panguiarchaeum genomes* and GCA_029856635.1 form a sister group to *Njordarchaeaceae* and branch within Asgard archaea**

137 *Assembling a set of vertically-evolving archaeal marker genes*
138
139 We used a range of phylogenetic analyses to investigate the evolutionary relationships
140 among Panguiarchaeales and Njordarchaeales, with Korarchaeota[26] and Asgard
141 archaea (officially referred to as Asgardarchaeota based on SeqCode[35] and as
142 Promethearchaeota[36] based on International Code of Nomenclature of Prokaryotes).
143 To begin with, we assembled a set of marker genes that appear to have evolved
144 vertically within Archaea and thus can be used to trace the species tree. Our starting
145 point was the set of genes that have been used to investigate archaeal phylogeny in
146 several previous analyses, including ribosomal proteins as well as genes with
147 alternative functions[4,37,38], totalling 120 distinct homologous Archaeal Clusters of
148 Orthologous Genes (arCOG) gene families. We then performed a series of analyses
149 to select the most reliable marker genes for concatenation, filtering out families that
150 have been affected by gene replacement transfers from Bacteria (9 families) that were
151 found in less than 60% of taxa (4 markers) or which have a complex history of gene
152 duplication, transfer and loss within Archaea, and so are unreliable markers for
153 inferring vertical evolution (11 families, including some subunits of DNA topoisomerase
154 - see Supplementary Information, **Supplementary Data 1-3**, **Supplementary Figures**
155 **1-8**). We ranked the remaining markers based on the degree to which they recovered
156 accepted monophyletic archaeal clades (see Methods/Supplementary information,
157 **Supplementary Data 3**) and selected the top 50% of these markers corresponding to
158 43 proteins (**Supplementary Data 3**) for phylogenetic analyses.
159
160 *Phylogenetic analyses that model site- and branch- compositional heterogeneity place*
161 *Panguiarchaeaceae and Njordarchaeaceae within Asgard archaea*
162
163 Previous analyses have suggested that Panguiarchaeales and/or Njordarchaeales
164 may branch either with Korarchaeota[31,32] or Asgard archaea[4] (**Figure 1A**). To
165 distinguish between these hypotheses (i.e. Korarchaeota versus Asgard archaea
166 topologies (**Figure 1A**)) and assess the taxonomic ranks of these clades, we
167 performed phylogenetic analyses on the 43-gene concatenation using distinct taxon
168 sets as well as relative evolutionary distance (RED) scaling to normalise the taxonomic
169 ranks[35,38] (see Methods and **Supplementary Figure 1** for method overview). All
170 analyses provided strong support for the monophyly of the previously proposed

171 Njordarchaeales and Panguiarchaeales clades, together forming a putative class-level
172 taxon, i.e. the Njordarchaeia. The Njordarchaeia comprises the Njordarchaeales order
173 with two family-level lineages, the *Panguiarchaeaceae* and *Njordarchaeaceae* (**Figure**
174 **1BC, Supplementary Figure 9,** See also Supplementary Information). Analyses
175 using the best-fitting substitution models robustly supported the placement of the
176 Njordarchaeia within the Asgard archaea (**Figure 1BC**, **Supplementary Figure 1**).
177 However, we did observe the Korarchaeota topology, i.e. the placement of all
178 Njordarchaeia with Korarchaeota, in several analyses using simpler substitution and
179 empirical distribution mixture (EDM) models (LG+G+F, LG+EDM0004LCLR+G+F,
180 LG+EDM0008LCLR+G+F, **Supplementary Figure 1** and **Supplementary Figure**
181 **10A-C**), consistent with the hypothesis that this inferred relationship might be a
182 phylogenetic artefact[4].
183
184 Current phylogenetic methods are not consummate, and even the best models are
185 mere approximations of the evolutionary processes that gave rise to the observed
186 data. However, previous work has identified two aspects of the evolutionary process
187 that are not well-captured by simple models, which have, in previous case studies, had
188 a substantial impact on phylogenetic inference. These are variations in amino acid
189 preferences across the alignment sites (due to varying protein biochemical
190 constraints) and across the tree branches (due to varying environmental pressures
191 and mutational biases).
192
193 *Modelling site-wise compositional heterogeneity.* Site-wise compositional variation
194 results from selective constraints on protein function: typically, only a small subset of
195 the twenty possible amino acids are permissible at a given position in the protein due
196 to the requirement for particular types of side chains to maintain protein structure and
197 function[39]. While being a pervasive feature of molecular sequence data, this site-wise
198 variation is not accounted for in simple substitution models such as LG+G, which
199 assume that all sites have the same equilibrium amino acid frequencies[40,41]. As they
200 do not account for site-specific selective constraints, simple models under-estimate
201 the probability of convergent evolution to the same amino acid in independent,
202 distantly-related lineages. This means they can sometimes mis-interpret instances of
203 molecular convergence as evidence that these taxa are closely related, an artefact
204 that is expected to be strongest at the most selectively-constrained positions. To

205 evaluate whether under-estimation of convergent evolution might drive the support for
206 the cladehood of Njordarchaeia and Korarchaeota in analyses with simpler models,
207 we binned amino acid sites by compositional constraint and investigated support for
208 the two trees from each class of sites (**Figure 1D**). This analysis revealed that the
209 support for the Korarchaeota tree in simpler models was greatest at the most
210 compositionally constrained sites, whereas with the best-fitting model, both
211 constrained and compositionally variable sites supported the Asgard archaea tree
212 (**Figure 1D**; **Supplementary Figures 1,10-13**). Additionally, under the best-fitting
213 model, the Korarchaeota tree is confidently rejected by the approximately unbiased
214 (AU) test ($p < 0.05$) (**Supplementary Data 4**). Recoding based on the KGB6 and SR4
215 scheme also recovered the Asgard archaea topology (**Supplementary Figures 14-**
216 **15**). These analyses indicate that failure to model site-wise compositional
217 heterogeneity contributes to the incorrect placement of Njordarchaeia with
218 Korarchaeota in analyses with simpler models.
219
220 *Modelling branch-wise compositional heterogeneity.* In addition to site-wise variation,
221 sequence composition can also vary among lineages with, for example, distantly-
222 related thermophiles converging on greater use of thermostable amino acids (Ile, Val,
223 Tyr, Trp, Arg, Glu, Leu (IVYWREL[42]) or charged amino acids: Asp, Glu, Lys, Arg
224 (DEKR)[43,44]). Njordarchaeia and Korarchaeota were both predicted to be
225 hyperthermophiles, and the placement of Njordarchaeia with Korarchaeota was
226 suggested to be a result of compositional biases[4]. In line with a previous study[4], a
227 comparison of proteome composition across Archaea indicated that they indeed have
228 similar amino acid compositions (**Supplementary Figure 16AB**). We subsequently
229 investigated whether compositional attraction between the two lineages might also
230 contribute to the recovery of the Korarchaeota tree under simple models. First, we
231 performed a site-filtering analysis. We ranked sites by compositional bias, iteratively
232 removing the sites that made the greatest individual contribution to across-branch
233 compositional heterogeneity (see Methods), and re-evaluated support for the
234 competing topologies (**Figure 1E**). Removal of the top 20% of most biased sites shifted
235 support from the Korarchaeota topology to the Asgard archaea topology under
236 LG+C60+G+F, while removal of biased sites under LG+EDM00256LCLR+G+F model
237 does not affect the support of the Asgard archaea topology, consistent with the
238 hypothesis that failure to model branch-wise compositional heterogeneity underlay

239 support for the Korarchaeota tree in the full dataset. Second, we identify RYPEIWKVL
240 and QSNTDGC, as overrepresented and underrepresented amino acids, respectively,
241 in the concatenation[42] (**Supplementary Figure 16CD**). Subsequently, we performed
242 phylogenetic analysis using a substitution model that accounts for changing
243 compositions across the tree, i.e. GFMix[45]. The LG+EDM00256LCLR+G+F+GFMix
244 (RYPEIWKVL/QSNTDGC and RYPEI/QSNT) and LG+C60+G+F+GFmix
245 (RYPEIWKVL/QSNTDGC and RYPEI/QSNT) model provided strong support for the
246 Asgard archaea tree over the Korarchaeota tree (**Supplementary Data 5**), again
247 consistent with the hypothesis that compositional attraction due to a shared
248 environmental adaptation contributes to the Korarchaeota tree in analyses with models
249 that do not account for branch heterogeneity[4].
250
251 *A close relationship between Njordarchaeia and Asgard archaea is supported by*
252 *shared eukaryotic signature proteins (ESPs) and protein presence-absence profiles*
253
254 The phylogenetic analyses described above provided strong support for the placement
255 of *Panguiarchaeaceae* within Njorarchaeales as a sister to *Njordarchaeaceae* within
256 the Asgard archaea (**Figure 1B-E**). Consistent with that placement, we observed that
257 *Panguiarchaeaceae* and *Njordarchaeaceae*[4] encode a number of eukaryotic signature
258 proteins (ESPs) shared with other Asgard archaea (**Supplementary Figure 17,**
259 **Supplementary Data 6**), based on analyses using the Asgard Cluster of Orthologues
260 (AsCOGs) database[3]. Of particular interest, we identified homologs of ribosomal
261 L28e/Mak16 (PF01778) in *Panguiarchaeaceae* MAGs, whose distribution in published
262 archaeal genomes is otherwise restricted to other Njordarchaeia, the Hodarchaeales,
263 and Wukongarchaeia[4]. Phylogenetic analysis of ribosomal L28e/Mak16 revealed a
264 topology consistent with the species tree **(Supplementary Figure 18**), with the
265 Njordarchaeia sequences branching sister to Wukongarchaeia and other
266 Heimdallarchaeia. Furthermore, *Panguiarchaeaceae* MAGs encode an actin homolog
267 that branches with the conserved lokiactin clade present in all asgard archaeal
268 lineages (**Supplementary Figures 19-20**). Homologs of proteins involved in the N-
269 glycosylation process were also identified in *Panguiarchaeaceae* as well as
270 *Njordarchaeaceae* MAGs[4], and the phylogenetic analyses support their affiliation with
271 Asgard archaeal homologs (**Supplementary Figures 21-22**, see Supplementary
272 Discussion for details). Phylogenetic analyses of Njordarchaeia SNF7 proteins also

support them as Asgard archaeal homologs (**Supplementary Figure 23**, see Supplementary Discussion for details).

Finally, protein presence-absence profiles revealed a pattern of transcription, translation and replication distinct from Korarchaeota (**Supplementary Figures 24-25**). Interestingly however, the profile for other COG categories was more similar to that of Korarchaeota than to other Asgard archaea (**Supplementary Figures 24-25**), likely reflecting a result of shared environmental adaptation and horizontal gene transfer.

Incomplete biosynthetic pathways indicate that *Panguiarchaeum* may depend on other organisms

Comparative genome analyses and the inference of metabolic potentials for Njordarchaeia MAGs (i.e. the *Panguiarchaeaceae* clade including *Panguiarchaeum* and GCA_029856635.1 and the *Njordarcheaceae* clade) suggest that these organisms are thermophilic fermentative heterotrophs that have varying levels of auxotrophies; that is, they cannot synthesise all organic compounds needed for growth (**Supplementary Figure 26**). Of note, members of the *Panguiarchaeum* show a reduced metabolic potential, in particular in purine and lipid biosynthesis pathways, compared to Njordarchaeacea and Korarchaeota (**Supplementary Figure 26-30;** see also Supplementary Discussion).

Catabolism and energy conservation. The catabolic potential of the *Panguiarchaeum* clade (i.e. upon divergence from GCA_029856635.1) is more limited than that of *Njordarchaeaceae* (**Figure 2**, **Supplementary Data 7-9, Supplementary Figure 26**). In particular, while both clades encode a nearly complete Embden-Meyerhof-Parnas (EMP) pathway, members of *Njordarchaeaceae* encode more genes for enzymes of the tricarboxylic acid cycle (TCA) and beta-oxidation pathway. Furthermore, members of both clades might have the potential to grow on organic substrates, which could be fermented to acetate by a putative acetate-CoA ligase (ACD) with concomitant production of ATP[46]. Additionally, we identified homologues of the small and large subunits of [NiFe]-hydrogenases belonging to group 3 and group 4 in *Panguiarchaeum*

307 MAGs. Phylogenetic analyses of the large subunit of the [NiFe] hydrogenase indicated
308 that panguiarchaeal homologs belong to subgroup 4g, forming a clade with
309 Odinarchaeia sequences (**Supplementary Figure 31A**). Group 4g [NiFe]-
310 hydrogenases of *Panguiarchaeum* encode a Nuo-L-like subunit, which might be
311 involved in sodium/ion translocation[47] (**Supplementary Figure 31B**). This suggests
312 that *Panguiarchaeum* may conserve energy by coupling the oxidation of ferredoxin to
313 the formation of H_2 to generate a proton gradient for ATP synthase. Members of
314 *Panguiarchaeum* also contain various group 3 [NiFe]-hydrogenases assigned to
315 subgroup 3c and 3b, which may be involved in the reoxidation of cofactors during
316 carbon oxidation and the production of H_2, a reversible process (**Supplementary**
317 **Figure 32, Supplementary Data 10**). Together, this indicates that representatives of
318 the *Panguiarchaeum* clade can conserve energy through substrate-level
319 phosphorylation as well as via the ATP synthase using a proton gradient generated by
320 the oxidation of ferredoxin similar to Odinarchaeia and Thermococci[14].
321 *Panguiarchaeum* representatives revealed a potential to harness electrons from
322 organic substrates such as amino acids, peptides, 2-ketoacids (e.g., pyruvate,
323 oxoglutarate, and indole-pyruvate), and aldehydes (**Supplementary Data 7**,
324 Supplementary Information).
325
326 *Anabolism*. The biosynthetic potential of members of the *Panguiarchaeum* clade
327 seemed more restricted than that of *Njordarchaeaceae* (**Supplementary Figure 26**).
328 For instance, all three *Panguiarchaeum* MAGs lack various genes encoding enzymes
329 for *de novo* purine biosynthesis (from phosphoribosyl pyrophosphate to inosine
330 monophosphate), such as glutamine phosphoribosylpyrophosphate amidotransferase
331 (PurF) and Phosphoribosylaminoimidazole (AIR) synthetase (PurM) indicating a
332 possible dependence on external purine sources[48]. Notably, *Panguiarchaeum* clade
333 MAGs do not encode enzymes of the mevalonate pathway (MVAP) for isoprenoid
334 biosynthesis and fatty acid biosynthesis, as indicated previously[32]. However, the
335 presence of CDP-archaeol synthase (CarS), phosphatidylglycerophosphate synthase
336 (PgsA), phosphatidylglycerophosphate synthase (PssA) and cardiolipin synthase
337 (Cls), suggests that *Panguiarchaeum* have the potential to attach polar head groups
338 to di-o-geranylgeranylglyceryl phosphate (DGGGP). The *Panguiarchaeum* MAGs
339 also encode a gene for digeranylgeranylglycerophospholipid reductase (GGR), which
340 could catalyse the hydrogenation/saturation of geranylgeranyl chains of DGGGP[49].

341 Phylogenetic analyses of CarS showed that the homologs of *Panguiarchaeum*
342 branched sister to Heimdallarchaeia with moderate support (UFBOOT/SH-aLRT:
343 88.8/83; **Supplementary Figure 33**). Similarly, although the PgsA homologs of
344 *Panguiarchaeum* branched sister to a clade comprising one Thermococcus and two
345 Heimdallarchaeia sequences with weak support (UFBOOT/SH-aLRT: 50.5/83), they
346 were nested within a cluster of predominantly Asgard archaeal sequences
347 (**Supplementary Figure 34**), Thus, these genes were likely inherited vertically from
348 their common ancestor with Heimdallarchaeia (**Supplementary Figure 33-35**).

349 ***Panguiarchaeum* has evolved by genome streamlining from a larger ancestor**
350 **with Heimdallarchaeia**

351 Reconciliation analyses have been used to infer the ancestral genome content of
352 various archaeal lineages[4,50–54], including Asgard archaea[4], shedding light on
353 ancestral gene repertoires and gene content evolution. To reconstruct the evolution of
354 *Panguiarchaeum*, we reconciled single gene trees of 6077 arCOG gene families with
355 the inferred species tree (from streamlined set 2 with 303 taxa, **Figure 1B**) to
356 determine the pattern and extent of reductive genome evolution in the
357 *Panguiarchaeum* clade and its sister lineages.

358 Our analysis suggests an ongoing decrease in gene content through time in the
359 evolution of Njordarchaeia, similar to a previous analysis[4], from around 1951 arCOG
360 families in the common ancestor with Wukong-Heimdallarchaeia to 911-1545 (mean:
361 1186.5) arCOG families among extant Njordarchaeia (**Supplementary Data 11**).
362 Based on the relationship between arCOG family members and the genome size of
363 extant taxa (Pearson's correlation coefficient: 0.89, P-value: 1.96x1e-110), we inferred
364 the genome size of the last common ancestor of Wukong-Njord-Heimdallarchaeia to
365 be 2.85 Mbp (95% confidence interval, 2.57 to 3.06, **Supplementary Data 11**). A
366 notable reduction in genome size is inferred at the branch leading to the Njordarchaeia
367 (from 0.53 to 0.84 Mbp; node id: 571-563) after divergence from their common
368 ancestor with Wukong-Heimdallarchaeia (**Figure 3A**, **Supplementary Data 11**). The
369 arCOG families lost along this branch include components of amino acid biosynthesis,
370 de novo purine biosynthesis (from ribose-5-phosphate to inosine monophosphate
371 (IMP)), Wood-Ljungdahl pathway genes and cofactors (methanofuran) biosynthesis
372 (**Figure 3B**, **Supplementary Data 12**). Additional genes, including those encoding

various enzymes involved in archaeal lipid biosynthesis and beta-oxidation, appear to have been lost along the branch leading to the *Panguiarchaeum* clade after its divergence from GCA_029856635.1 and the common ancestor with *Njordarchaeaceae*, which resulted in an inferred genome reduction of 0.13 Mbp to 0.17 Mbp. *Panguiarchaeum* and *Njordarchaeaceae* were inferred to encode fewer ESPs than the other Asgard archaea (**Supplementary Figure 17**), and the loss on the *Njordarchaeaceae* and *Panguiarchaeaceae* stem is predicted to include the reduction in ESP repertoire as well. For instance, we inferred the loss of a Roadblock/LC7 domain-containing protein (arCOG02605_01) on the *Njordarchaeaceae* branch; the loss of Ras family GTPase (arCOG05343_01) and Yip1 domain family (arCOG02054_01) on the *Panguiarchaeaceae* branch; the loss of VPS4 (arCOG01307_01) on the *Panguiarchaeum* branch (i.e. upon divergence from GCA_029856635.1) (**Supplementary Data 12**); and the absence of Snf7 family proteins in the extant *Panguiarchaeum* MAGs indicated their loss on the *Panguiarchaeum* branch (**Supplementary Figure 23**).

Njordarchaeia encodes reverse gyrase, a topoisomerase typically encoded by hyperthermophiles[55]. Consistent with previous work[4] and proposal of a thermophilic ancestry of Asgard archaea[56], we inferred the presence of reverse gyrase (Presence probability, PP = 0.96) in the last common ancestor of Asgard archaea. This gene family was, however, subsequently lost in Heimdallarchaeia (**Supplementary Data 12**) but not in Wukongarchaeia and Njordarchaeia (See Supplementary Information). Interestingly, phylogenetic and reconciliation analyses suggest a later re-acquisition of a reverse gyrase in the *Panguiarchaeum* clade (**Supplementary Figure 36,37**; See Supplementary Information). The MAG/genome size of putative hyperthermophilic Asgard archaea encoding reverse gyrase (mean: 2.08 Mbp, n=18) is significantly smaller than that of Asgard archaea representatives without reverse gyrase (mean: 3.54 Mbp, n=45) (p=5.669x10-7, two-tailed Wilcoxon rank sum test). Genome reduction has previously been reported in cases of thermophilic adaptation, perhaps due to a selective advantage for smaller cell volumes (and so genome) at high temperatures [57,58]. We, therefore, hypothesise that the initial reductive evolution of Njordarchaeia might have been associated with their adaptation to high-temperature environments. The further genome reduction in *Panguiarchaeum*, including the loss of

405 essential genes for biosynthetic pathways, might be associated with a transition to a
406 host-associated lifestyle[59].

407 ***Panguiarchaeum* are associated with Thermoproteota taxa**

408

409 To investigate the inferred dependence of *Panguiarchaeaum* on other community
410 members, we assessed the extent to which members of this group co-occur with other
411 microbes using network inferences. First, we tested the inference of our co-occurrence
412 networks with the known host-symbiont pair *Nanoarchaeum equitans* and *Ignicoccus*
413 *hospitalis*[60], and the proposed symbiosis between *Huberarchaeum crystalense* and
414 *Altiarchaeum hamiconexum*[61]. We did not detect any associations between the former,
415 mainly due to insufficient sampling of *N. equitans* (**Supplementary Figure 38-39,** See
416 Supplementary Information). However, for *Huberarchaeum-Altiarchaeum,* we found a
417 high correlation (Spearman's rho = 0.931) and a robust detection frequency (100%) of
418 this suggested interaction starting at a sample size of 40 (**Supplementary Figure 39F**;
419 see Methods). Next, we identified 120 samples containing taxa annotated as family
420 *Panguiarchaeaceae* in GTDB[38], a lineage equivalent to the order Njordarchaeales
421 proposed in this manuscript, from 248,559 NCBI metagenomes community profiles in
422 the 'Sandpiper' website (https://sandpiper.qut.edu.au). Community profiles of these
423 120 samples were subjected to various data transformations (see methods) and
424 network inferences. Across all original networks, i.e. independent of filtering criteria
425 applied before network inference, Njordarchaeales taxa were associated with
426 Thermoproteota taxa. The frequency of association between these taxa was up to ~4
427 times higher than that between Njordarchaeales and any other taxa or random number
428 controls. On phylum, class, and order level Njordarchaeales were associated with
429 Thermoproteota (73 out of 256), Thermoprotei_A (33 of 73), and Sulfolobales (23 of
430 33) (**Supplementary Data 13-14**), respectively, in agreement with abundance-based
431 co-occurrence patterns (**Figure 4)**. Within the order Njordarchaeales, there was only
432 one OTU (OTU_804; **Supplementary Data 14**) assigned to the genus
433 *Panguiarchaeum*, that found association partners in the network. This genus had 29
434 associations in total, whereby the taxonomic annotation of its partners was similar to
435 the associations detected for the order, i.e. Thermoproteota (15 of 29),
436 Thermoprotei_A (9 of 15), and Sulfolobales (4 of 9), and *Desulfurococcaceae* (3 of 4)
437 (**Supplementary Data 14**). These results indicate that members of the

438 Njordarchaeales, in particular the genus *Panguiarchaeum*, might interact with
439 members of the phylum Thermoproteota, in particular with Sulfolobales, an order
440 comprising hosts of thermophilic Nanoarchaeota[60,62–67].
441
442
443 **Discussion**
444
445 Reconstructing the tree that best depicts the evolutionary relationship between major
446 archaeal lineages is important for our understanding of archaeal genome evolution
447 and the evolution of archaeal symbiosis and eukaryogenesis. Our comprehensive
448 phylogenetic and comparative genomic analyses reveal that *Panguiarchaeum*
449 together with GCA_029856635.1 forms a family within the Njordarchaeia as part of the
450 Asgard archaea. In contrast, the placement of this clade as a sister group to
451 Korarchaeota appears to be due to phylogenetic artefacts[4], including the use of
452 problematic marker gene families in concatenations and insufficient modelling of
453 compositional biases and site heterogeneity. These results emphasise the importance
454 of evolutionary model fit in analyses of prokaryotic evolution, particularly for datasets
455 containing pervasive compositional heterogeneity.
456
457 The placement of *Panguiarchaeum* within the Asgard archaea is supported by the
458 presence of various ESPs shared with *Njordarchaeaceae*. However, metabolic
459 reconstructions reveal that members of *Panguiarchaeum* have various auxotrophies
460 and encode fewer ESPs than other Asgards. This is indicative of genome reduction
461 potentially associated with a host-associated lifestyle, with co-occurrence analyses
462 pointing towards an archaeal host. Notably, *Ca*. P. syntrophicum, the first cultivated
463 representative of the Asgard archaea (i.e. a member of the
464 Lokiarchaeia/Prometheoarchaeia), appears to syntrophically grow on amino acids with
465 partner organisms such as *Halodesulfovibrio* and *Methanogenium*[13]. However, this
466 symbiotic interaction does not seem to be obligate and has not resulted in reductive
467 genome evolution. Our inferences of the energy metabolism of *Panguiarchaeum,* and
468 perhaps Njordarchaeaceae as well, which seems to resemble that of Odinarchaeia
469 and *Thermococcus*, indicates the potential for syntrophic growth[14,47,68,69]. For example,
470 *Panguiarchaeum* encodes 3 different [NiFe]-hydrogenases, including a putative
471 membrane-bound group 4g [NiFe]-hydrogenase, and might use electrons harnessed

472 from small organic acids/simple carbohydrates to produce H_2. While speculative in the
473 absence of experimental data, it may be hypothesised that the hydrogen could support
474 the growth of a H_2-ultilising archaeon, perhaps, in return for essential cellular building
475 blocks such as lipids and vitamins[13,69–71].
476
477 Likewise, it is interesting to note that our ancestral reconstructions infer a complex
478 ancestor of Asgard archaea as was observed previously[4]. Further, our analyses
479 predict that the Asgard archaeal ancestor has encoded a Wood-Lundgal pathway
480 (WLP) (**Figure 3**, **Supplementary Data 12**), i.e. a pathway that can serve as electron
481 sink supporting both autotrophic and heterotrophic growth modes[72–74]. However, this
482 pathway seems to have been lost in various members of the Asgard archaea[3,4,14]
483 (**Supplementary Data 9, 12**); specifically, our analyses indicate that genes for
484 proteins involved in the WLP were lost on the branch leading to the Njordarchaeia after
485 their split from Heimdall- and Wukongarchaeia (**Figure 3**). The absence of the WLP
486 as electron sink may have strengthened the dependency of members of this group on
487 syntrophic H_2-ultilising partners, such as Thermoproteota[75–77], that could serve as
488 external electron sinks. It is tempting to speculate that such a syntrophic lifestyle may
489 have permitted the loss of genes involved in various biosynthetic pathways in the
490 ancestor of the *Panguiarchaeum* clade (**Figure 2,3**), increasing the dependency of its
491 members on partner organisms.
492
493 This scenario on the evolution of a potentially obligate symbiont lineage among the
494 Asgard archaea points to interesting parallels with the evolutionary trajectory leading
495 to the origin of eukaryotes. Many current models on eukaryogenesis hypothesise
496 syntrophy as a key driver for the establishment of an intricate symbiotic relationship
497 between the archaeal ancestor of eukaryotes[7,13–15,78], i.e. a likely sister lineage of the
498 Hodarchaeales[4,24], and the alphaproteobacterial partner that evolved into the
499 mitochondrion[17,18,22,45,79].
500
501 Similar to *Panguiarchaeum*, the eukaryotic host lineage also appears to have lost
502 ancestral archaeal genes during eukaryogenesis, i.e. the transition from the First to
503 the Last Eukaryotic Common Ancestor. For example, glycolysis and iron-sulphur
504 cluster biosynthesis in eukaryotes are implemented by pathways with patchwork
505 ancestries that include both archaeal, alphaproteobacterial and other bacterial

contributions[80,81], while the archaeal TCA cycle was replaced by that of the mitochondrion in eukaryotes[80]. Similarly, the evolution of *Panguiarchaeum* was shaped by the loss of genes involved in the TCA cycle and several biosynthetic pathways. However, reductive evolution in *Panguiarchaeum* has also led to the loss of several genes encoding ESPs found in its close relatives. In contrast, the eukaryotic lineage expanded an already large protein repertoire and its ESPs via gene duplication and genome expansion, as well as acquisitions from mitochondrial and other bacterial genes along its stem[7,82–85]. Recent analyses have demonstrated that Asgard archaeal genomes often contain substantial proportions of horizontally-acquired genes of bacterial origin[5], and we hypothesise that part of the difference in evolutionary trajectories of *Panguiarchaeum* and eukaryotes may result from the nature of the symbiotic partner, either archaeal (*Panguiarchaeum*) or bacterial (eukaryotes) as well as environmental parameters (e.g. temperature). To address these hypotheses, prospective efforts would benefit from a better sampling of symbiotic members of the Asgard archaea including both syntrophic and/or genome-reduced representatives, combined with their co-cultivation with partner organisms and physiological characterisation.

540 **Materials and Methods**

541 **Marker gene inspection and ranking analysis**

542 **Marker gene inspection:** To accurately place Njordarchaeales and Panguiarchaeals
543 in the archaeal tree of life, we first downloaded the NM57 A64 dataset from Eme et
544 al., 2023[4] and assigned a cluster of orthologs (COG) family to each sequence using
545 hmmsearch[86] v3.3.2 (settings: -E 1e-5). The best hit for each protein sequence was
546 selected based on the lowest e-value and highest score, and the COG families
547 representing the majority of sequences in each marker gene were selected to annotate
548 each marker. These COG families were then compared to 185 marker gene families
549 used in four other previous studies[24,37] to create a non-redundant marker set
550 comprising 112 unique COG families for further inspection. Potential orthologous
551 sequences assigned to these 112 unique families were identified in 966 archaeal and
552 1325 bacterial genomes/MAGs using hmmsearch v3.3.2 (settings: -E 1e-5) and used
553 for subsequent phylogenetic analyses. Since gene fission of DNA-directed RNA
554 polymerase subunits A and B were reported in some archaea[87], we used a custom
555 HMM profile of COG0085 and COG0086 to detect "split" subunits and concatenate
556 these subunits before alignment (https://doi.org/10.5281/zenodo.14777939).
557
558 We then aligned sequences assigned to these 112 marker gene families using mafft
559 v7.453[88] and removed poorly aligned amino acid sites with BMGE 1.12[89] (settings: -m
560 BLOSUM30 -h 0.55). The maximum likelihood trees for these marker genes were
561 inferred using IQ-tree v2.1.2[90] (-m LG+G -B 1000). We decorated the tips of these
562 marker gene trees with annotations from KEGG orthology[91], PFAM[92], COG[93] and
563 archeal COG (arCOG)[94] information. We manually inspected all single protein
564 phylogenies to identify marker gene families that did not meet archaeal monophyly
565 and excluded those families from subsequent analyses. Furthermore, we identified
566 and removed paralogues and sequences that experienced horizontal gene transfer
567 (HGT) events from bacteria to archaea. In total, this yielded 100 COG marker gene
568 families.
569
570 We performed a second round of inspection and curation based on corresponding
571 arCOG families of these 100 COG families except COG0085 and COG0086 (See
572 above). In this round, we only included the marker gene families if homologs were
573 present in at least 60% of archaeal taxa and were duplicated in less than 20% of the

574 genomes of major archaeal lineages (e.g., Asgard archaea, Thermoproteota and
575 Halobacteriota), which resulted in 90 unique families with arCOG identifiers.
576 Sequences were realigned with mafft-linsi v7.453, and poorly aligned sites were
577 removed using BMGE as described above. The initial ML trees were inferred using IQ-
578 tree v 2.1.2[90] (-m LG+G -B 1000). These trees were then used as guide tree to perform
579 phylogenetic inference based on the best-fitting model [95] (settings: -mset LG -madd
580 LG+C10,LG+C20,LG+C30,LG+C40,LG+C50,LG+C60,LG+C10+R+F,LG+C20+R+F,
581 LG+C30+R+F,LG+C40+R+F,LG+C50+R+F,LG+C60+R+F --score-diff all) with 1000
582 ultrafast bootstraps. We inspected the single gene trees to identify and remove
583 sequences which seem to represent paralogues and/or were affected by long-branch
584 attraction (LBA) artefacts or HGT.
585
586 **Ranking analysis:** We used a previously developed marker gene ranking procedure
587 to rank each of the 90 markers based on the extent to which they recovered
588 established archaeal phylum- or order-level lineages[37]. In brief, the number of splits
589 (i.e., the occurrence of a certain taxon failing to group within its expected taxonomic
590 clade, **Supplementary Data 15**) was counted across all bootstrap trees for each
591 marker gene. We then defined the highest ranking (top 50%, 45 markers) marker sets
592 based on the number of splits per phylogenetic cluster and the total number of splits
593 normalised by the number of genomes within each tree.
594
595 **Concatenation, recoding, site filtration and species tree inference**
596 **Concatenated gene trees using taxon set 1: i.e. 966 taxa:** The top 50% marker
597 gene sequences were aligned with mafft-linsi v7.453[88] and poorly aligned amino acid
598 sites were removed with BMGE 1.12[89] (settings: -m BLOSUM30 -h 0.55). Genomes
599 with more than 50% gaps in the alignment were further removed. The trimmed
600 sequences of these 45 markers were then subsequently concatenated into a
601 supermatrix comprising 966 taxa. Two additional supermatrixes were constructed by
602 either excluding DNA topoisomerase VI subunit A (arCOG04143) alone or together
603 with DNA topoisomerase VI subunit B (arCOG01165) for the following reasons (See
604 also Supplementary Information):
605 1. The mixing of orthologues of arCOG04143 provides spurious support for the
606 relationship between Njordarchaeia and Asgard archaea.

607 2. The interaction between DNA topoisomerase VI subunits might lead to co-
608 evolution of some amino acid residues.
609
610 Concatenated gene trees were inferred using IQ-Tree v2.1.2[90] under LG+C60+G+F
611 model with posterior mean sites frequency approximation (PMSF)[96] with guide trees
612 inferred under LG+G+F model.
613
614 **Concatenated gene trees using taxon set 2: i.e. 303 taxa:** To resolve the
615 relationship between Njordarchaeales, Panguiarchaeales, Korarchaeota and Asgard
616 archaea, we next excluded divergent sequences from DPANN MAGs/genomes and
617 downsampled the Thermoproteota (previously TACK), and Halobacteriota,
618 Thermoplasmatota, Hydrothermarchaeota, Methanobacteriota, Methanobacteriota_A
619 and Methanobacteriota_B lineages (previously "Euryarchaea") based on the predicted
620 genome completeness and contamination (completeness-5 X contamination)[97] to
621 create a dataset of 303 taxa by one-per-Thermoproteota/Asgardarchaeota-family and
622 one-per-"Euryarchaea"-order citeria. Marker gene sequences were realigned with
623 mafft-linsi v7.453[88] and trimmed with BMGE 1.12[89] (settings: -m BLOSUM30 -h 0.55).
624 We concatenated three supermatrices by including all top 50% ranked markers and
625 excluding arCOG04143 alone or with arCOG01165. Trees were inferred under the
626 LG+G+F and LG+C60+G+F model, as well as the LG+EDM0256LCLR+G+F model
627 using IQ-Tree v2.1.2[90] and IQ-Tree v2.2.2.7[90] respectively (settings: -B 1000 -alrt
628 1000, see also 71 taxa). Focusing on the supermatrix of 303 archaeal taxa, excluding
629 DNA topoisomerase subunits, we progressively removed 10% to 50% of fastest
630 evolving sites according to the empirical Bayesian rate under LG+C60+G+F model
631 (settings –rate), and 5%, 10% to 50% most heterogeneous sites using
632 Alignment_pruner.pl (https://github.com/novigit/davinciCode/blob/master/perl). Trees
633 were inferred under either LG+C60+G+F or LG+EDM0256LCLR+G+F models
634 (settings: -B 1000 -alrt 1000).
635
636 **Concatenated gene trees using taxon set 3: i.e. 71 taxa:** A smaller dataset of 71
637 taxa was created based on the 303 taxa set (set 2) for Bayesian phylogenetic
638 analyses. In particular, we kept one genome of "Euryarchaea" for each GTDB R207
639 class as an outgroup and one genome of thermoproteotal (TACK) lineages for each
640 GTDB R207 order considering predicted genome completeness and contamination

641 (completeness-5xcontamniation)[97]. Marker gene sequences were realigned with
642 mafft-linsi v7.453[88] and trimmed with BMGE 1.12[89] (settings: -m BLOSUM30 -h 0.55).
643 We concatenated only one supermatrix without both DNA topoisomerase subunits
644 (arCOG04143 and arCOG01165). ML trees were inferred under LG+G+F and
645 LG+C60+G+F model using IQ-Tree v2.1.2[90] (settings: -B 1000 -alrt 1000).
646
647 Additionally, we performed a comprehensive phylogenetic analysis using a recently
648 developed CAT-PMSF pipeline[98]. In particular, we obtained three different guide trees
649 inferred under LG+G+F, LG+C60+G+F and GTR+G+C60SR4 (see Recoding section
650 below), representing different hypotheses regarding the relationship of Njordarchaeia
651 to Korarchaeota and Asgard archaea, namely, Korarchaeota topology: Njordarchaeia-
652 sister-to-Korarchaeota (LG+G+F); Asgard archaea topology: Njordarchaeia-sister-to-
653 Wukong-Heimdallarchaeia (GTR+G+C60SR4) and Asgard archaea sister topology:
654 Njordarchaeia-sister-to-Asgard-Archaea (LG+C60+G+F). We then used these 3
655 different starting topologies in a subsequent bayesian analysis with the CAT model in
656 combination with either Poisson or LG exchangeabilities and a 4 discrete categories
657 gamma rate model. We ran four different Markov chains for each topology
658 representing different hypotheses under the Poisson+CAT+G model in Phylobayes v
659 1.8[99]. Additionally, four different Markov chains were run for both Asgard archaea and
660 Korarchaeota topology under the LG+CAT+G model. These chains were run until the
661 effective sample size of all parameters was above 100 and/or the relative divergence
662 of all parameters was below 0.3, or visual inspection with Tracer 1.7[100] indicated
663 convergence (**Supplementary Data 16**). Next, we extracted posterior mean site-
664 specific stationary distributions of amino acids for each chain using readpb_mpi[99],
665 which were converted to site-specific frequency profile using the python script convert-
666 site-dist.py (https://github.com/drenal/cat-pmsf-paper/blob/main/scripts/convert-site-
667 dists.py)[98]. The site-specific frequency profile was then used to infer ML trees using
668 IQ-Tree v2.1.2 (settings: -fs .sitefreq -B 1000 -alrt 1000).
669
670 We next applied the EDCluster algorithm[101] on one of the LG+CAT+G4 chains that
671 yielded the highest likelihood under LG+CAT-PMSF model to infer empirical
672 distribution mixture models (EDM) with various components (8, 16, 32, 64, 128 and
673 256) using log centred log-ratio transformation (LCLR). ML trees were then inferred
674 under different components of EDM models using IQ-Tree v2.2.2.7 (e.g., settings: -m

675 LG+EDM0256LCLR+G+F -B 1000 -alrt 1000) and universal distribution mixture
676 models (UDM) with 128, 256 and 512 components[101].
677
678 **Recoding**: All supermatrices comprising 303 and 71 taxa were further recoded into
679 different states using Susko-Roger 4 states[102] (SR4), Susko-Roger 6 states[102] (SR6),
680 Dayhoff 6 states(Dayhoff6), and Kosiol-Goldman-Buttimore 6 states[103] (KGB6)
681 recoding schemes, to explore the effect of recoding on mitigating compositional
682 biases. Respective ML trees were inferred under GTR+C60+SR4, GTR+C60+SR6,
683 GTR+C60+Dayhoff6 and GTR+C60+KGB6 models
684 (https://github.com/xgrau/recoded-mixture-models) (settings: -B 1000 -alrt 1000).
685
686 **Compositional constrained analyses**
687 Constrained searches were performed for Asgard archaea (-sister) and Korarchaeota
688 topology based on the unconstrained ML tree under the same model using IQ-Tree
689 v2.1.2[90] via the -g and -wslr flags. AU tests were performed using consel v 1.20[104] with
690 10000 bootstrap replicates. Site-wise log-likelihood differences were calculated for
691 Asgard archaea and Korarchaeota topology under LG+G+F, LG+C60+G+F,
692 Poisson+CAT-PMSF and LG+CAT-PMSF models to investigate the attraction of
693 Njordarchaeia towards Korarchaeota. The effective amino acids (K_{eff}) values of
694 mixture models and site-specific frequency profile were calculated using the convert-
695 site-dists-to-k_eff.py script[98,101]. The site-wise log-likelihood differences were then
696 grouped into different bins according to partitions, evolving rates and the effective
697 amino acids.
698
699 **RED values scaling and classification of ranks**
700 Relative evolutionary distance (RED) metric on the species trees was calculated with
701 methods described in the PhyloRank v1.1.12 tool[105]. Class, family and order rank
702 nodes in the 303 archaeal sets were extracted from our phylogenetic tree based on
703 LG+EDM00256LCLR+G+F model, as well as, from GTDB R220 reference tree[38].
704 Linear regression analysis of the RED values of our phylogenetic tree and GTDB R220
705 tree was performed using the Scipy package. The taxonomic boundaries used in
706 GTDB R220 were scaled onto our tree by the fitted slope and intercept.
707
708 **Amino acid composition**

709 We used a custom Python script to estimate the relative frequency of each amino acid
710 in our 71 taxa dataset for the concatenation without DNA topoisomerase subunits
711 (https://doi.org/10.5281/zenodo.14777939). A principal component analysis (PCA)
712 was performed using Scikit-learn[106], and two axes were plotted along with
713 eigenvectors using Seaborn[107]. We identified RYPEI(WKVL)/QSNT(DGC) as
714 thermophilic enriched/depleted amino acid residues in the concatenation using a
715 binomial test[51] and modified the GFmix[45] model to account for the variation of the ratio
716 RYPEI(WKVL)/QSNT(DGC) at every branch. The likelihood of different tree topologies
717 under the GFmix-RYPEI(WKVL)/QSNT(DGC) model was then calculated with
718 LG+C60+G+F and LG+EDM0256LCLR+G+F, with weights of mixtures model, branch
719 length, and alpha shape parameters estimated using IQ-Tree v2.1.2[90].
720
721
722 **Gene calling and Annotations**
723 We annotated 966 archaeal genomes/MAGs using an in-house annotation pipeline
724 (https://doi.org/10.5281/zenodo.14777939). In brief: Gene calling was performed
725 using Prokka[108] (v1.14.6, settings: –kingdom Archaea –addgenes –increment 10 –
726 compliant –centre UU –norrna –notrna). The generated protein files were searched
727 against COG database[93] (NCBI_COGs_Oct2020.hmm), arCOG database[93]
728 (All_Arcogs_2018.hmm), PFAM[92] (Release 34.0), TIGRFAM (Release 15.0), KEGG
729 orthology[91] (KO) profile (downloaded April, 2019), the Carbohydrate-Active enZymes
730 (CAZy) database[109] (downloaded from dbCAN2 in September 2019), the Transporter
731 Classification Database[110] (TCDB; downloaded in November 2018), the hydrogenase
732 database[111] (HydDB; downloaded in November 2018), the MERPOS database[112]
733 (Release 12.4) and NCBI_nonredudant database (NCBI_nr; downloaded in November
734 2018). Additionally, we used Interproscan version 5.61.93.0 to scan for protein
735 domains[113] (settings: --iprlookup --goterms).
736
737 COG (settings: -E 1e-5), arCOG (settings: -E 1e-5), PFAM (settings: -E 1e-5),
738 TIGRFAM (settings: -E 1e-5), KO (settings: -E 1e-5), and CAZy (settings: -E 1e-10)
739 identifiers were assigned using hmmsearch v3.3.2. The TCDB, HydDB and MERPOS
740 database were searched using BLASTp v2.12.0[114] (settings: -outfmt 6, -evalue 1e-20),
741 and NCBI_nr database was searched using DIAMOND v2.0.6[115] (settings: –more-
742 sensitive –e-value 1e-5 –seq 50 –no-self-hits –taxonmap prot.accession2taxid.gz).

743 Asgard COG[3] (asCOG) database was searched with psi-blast v2.12.0 (settings: -
744 show_gis -outfmt 6 -dbsize 100000000 -comp_based_stats F -seg no -evalue 1e-20).
745 The best hit of each protein for all other database searches was selected based on
746 the lowest e-value and highest score, and results were summarised for
747 Njordarchaeales MAGs. Additionally, 1325 bacterial genomes/MAGs and 137
748 eukaryotic genomes and/or largely complete transcriptomes (Supplementary Data 17)
749 were searched against COG, arCOG, PFAM and KO databases using the same
750 pipeline. Njordarchaeia MAGs annotation results were summarised in Supplementary
751 Data 6-10 and 18-20.
752
753 **Metabolic comparison**
754 Metabolic comparisons were based on the annotation results described above. We
755 counted the occurrence of each gene across each MAG/genome and reported the
756 presence and absence pattern using Pandas[116] in Python v3.9.0[117]. This data was
757 then summarised per GTDB R207 class level whenever possible. The Occurrence of
758 ESPs was compared using the results from asCOG searches. The presence and
759 absence pattern of genes across each database (COG_DB and ARCOG_DB) was
760 analysed with PCA and t-distributed Stochastic Neighbor Embedding (t-SNE) using
761 scikit-learn[106].
762
763
764 **Single gene tree inferences for ancestral reconstruction**
765 The predicted protein sequence of 303 archaeal genomes/MAGs was searched
766 against the arCOG database using hmmsearch v3.3.2[118] (settings: --tblout –domtblout
767 –notextw). We used the best hits (the lowest evalue and highest bitscore, cutoff: e-
768 value < 1E-3) to identify and split potential fused proteins. Specifically, we first
769 subtracted the position of the first domain hit from the full protein using bedtools
770 subtract[119] (v2.26.0). Secondly, we investigated the presence of a secondary domain
771 hit that is assigned to a different arCOGs than the first domain. Subsequently, we
772 repeated these two steps until all the proteins were investigated for up to four domains.
773 Finally, the individual domains from the original protein were extracted using the
774 positional domain information from hmmsearch output using bedtools getfasta
775 function. To keep track of unsplit and split domains from the original protein, we

assigned ‘a0’ notation for unsplit proteins and ‘a1’ to ‘a4’ for split domains (primarily to quarterly) in the generated trees.

We then combined all protein sequences and removed sequences with ambiguous amino acids using a custom Python script (remove_seq_with_specific_char.py, https://doi.org/10.5281/zenodo.14777939). To improve the phylogenetic signal in the single gene trees, we carefully performed three rounds of alignments before the final phylogenetic inference. First, we aligned sequences using mafft or mafft-linsi, for sequences with greater than or less than 1000 sequences, respectively. Alignments were then trimmed using trimAl 1.2rev59[120] with –gappyout flag. Trimmed sequences with greater than or equal to 50% gaps were removed using a custom Python script (faa_drop.py, https://doi.org/10.5281/zenodo.14777939). Second, we realigned the arCOG gene families (with more than or equal to 4 sequences) using MAFFT as described above and trimmed the alignment with BMGE (v1.12; settings: -m BLOSUM30 -b 2 -h 0.55) [89]. We then inferred a phylogenetic tree using IQ-Tree v2.1.2 (settings: -m LG+G)[90]. To alleviate the effect of long-branching sequences on phylogenetic inferences, we identified sequences/clusters of sequences on long branches of the inferred tree using a custom script (cut_gene_tree.py, https://doi.org/10.5281/zenodo.14777939) with a branch cutoff of two [121]. Single sequences were discarded, and clusters with more than four sequences were separated into different gene families. We appended the notation “_ [0-3]” to keep track of the separation of sequences into a new cluster. Finally, we aligned and trimmed the sequences as described in the second step. For each family, we selected the best-fit model using the model test with IQ-Tree v2.1.2 (settings: -m MF -mset LG -madd LG+C10, LG+C10+G, LG+C10+R, LG+C10+F, LG+C10+R+F, LG+C10+G+F, LG+C20, LG+C20+G, LG+C20+F, LG+C20+G+F, LG+C20+R, LG+C20+R+F, LG+C30, LG+C30+G, LG+C30+R, LG+C30+F, LG+C30+R+F, LG+C30+G+F --score-diff all -T 2). For the selected best-fitting model with profile mixtures, we used the posterior mean site frequency (PMSF) approximation[96] to infer the bootstrap tree distribution (settings: -s trimmed_aln.faa -m BEST-FITTING MODEL -T 2 -wbtl -B 1000 -pers 0.2 -nstop 500) using a guide tree inferred from LG+G model (settings: -T 2)[90]. For the best-fitting non-mixture models, bootstrap tree distribution was inferred using IQ-Tree v2.1.2 with the settings: -T 1 -wbtl -B 1000 -pers 0.2 -nstop 500. In total, we

809 generated 6077 single gene tree distributions in the analysis for ancestral
810 reconciliations.
811
812 **Gene tree and species tree reconciliations**
813 We used the amalgamated likelihood estimation (ALE, v1.0) approach to reconcile
814 single gene trees against the species tree. First, we generated the ALE objects from
815 the bootstrap trees that approximate tree uncertainty using ALEobserve. Using
816 genome completeness estimation from CheckM v2[122] and ALEml_undated[123]
817 program, we first reconciled the gene trees against species tree using the default
818 parameters, which assumes origination at any node of the species tree was uniform.
819 Based on the reconciliation results from this, we next used an additional reconciliation
820 approach described by Coleman et al., 2021[124] that assumes the origination
821 probabilities at the root of the species tree (O_R) are different to those for all other
822 internal nodes. We first inferred the O_R for each of the 21 arCOG functional
823 categories by maximising the total reconciliation likelihood over all gene families in that
824 category using two Python scripts (setup_OR_estimation.py and
825 O_R_optimisation.py, https://doi.org/10.5281/zenodo.14777939). Subsequently, we
826 used the category-specific probabilities for origination at the root to reconcile the single
827 gene trees and species trees and estimated the presence probabilities/copies of each
828 family in the internal nodes of the tree. The presence probabilities of both approaches
829 for each family for the nodes of interest are provided in **Supplementary Data 12**.
830 Ancestral genome size was predicted using locally weighted smoothing regression
831 based on the relationship between the genome size of extant taxa and their arCOG
832 gene family number. The uncertainty of the prediction of genome size was taken into
833 account using the bootstrap resampling approach (1000 times).
834
835 **Phylogenetic analyses of DNA topoisomerase subunit A and DNA**
836 **topoisomerase B**
837 The homologs of DNA topoisomerase subunit A and B, i.e. COG1389 and COG1697,
838 were identified and retrieved from 966 archaeal, 1325 bacterial and 137 eukaryotic
839 genomes/MAGs or largely complete transcriptomes. Sequences of each gene family
840 were first aligned with mafft-linsi v7.453, and poorly aligned alignment sites were
841 trimmed with BMGE v1.2[89] (Settings: -m BLOSUM30 -h 0.55). The inference of
842 phylogenetic trees for individual maker gene families was conducted as follows: an

843 initial ML tree was inferred under LG+G with IQ-Tree v 2.1.2[90]. Single gene trees were
844 manually inspected to remove putative LBA artefacts, and the remaining sequences
845 for each gene were realigned and trimmed as described above; finally, an ML tree was
846 inferred for each gene under LG+C60+G+F+PMSF [96] model (settings: -B 1000 -alrt
847 1000) using guide tree inferred from LG+F+G.
848
849 **Phylogenetic analyses of Actin homologs**
850 The actin homologs were identified from 966 archaeal, 1325 bacterial and 137
851 eukaryotic genomes/MAGs or largely complete transcriptomes using IPR00400. The
852 cell shape determining protein MreB homologs, a distant homolog to actin, in 966
853 archaeal genomes were identified using COG1077 and dereplicated using cd-hit v4.7
854 (-c 0.75)[125]. These sequences were combined and aligned with mafft-linsi v7.453[88].
855 Poorly aligned sequences were trimmed using BMGE[89] (settings: -m BLOSUM30 -h
856 0.55). An initial ML tree was inferred under LG+G using IQ-Tree v 2.1.2[90], and
857 sequences showing potential instances of LBA artefacts were removed; the remaining
858 sequences were then realigned and trimmed as described above; finally, an ML tree
859 was then inferred under LG+C60+G+F+PMSF model (settings: -B 1000 -alrt 1000)
860 using a guide tree inferred from LG+F+G. 20% of the most heterogeneous sites were
861 removed using alignment_pruner.pl and an ML tree was inferred in the same way
862 under LG+C60+G+F+PMSF using a guide tree inferred from LG+F+G
863
864 **Phylogenetic analyses of ribosomal protein L28**
865 The ribosomal protein L28/Mak16 homologs (i.e. PF01778) were retrieved from 966
866 archaeal, 1325 bacterial and 137 eukaryotic genomes/MAGs or largely complete
867 transcriptomes (evalue: 0.0001). Sequences were first aligned with mafft-linsi v7.453,
868 and gappy columns were removed using trimal 1.2rev59[120] with -gappyout flag. An
869 initial ML tree was inferred under LG+G using IQ-Tree v2.1.2[90], and sequences
870 showing potential instances of LBA artefacts were removed; remaining sequences
871 were then realigned and trimmed as described above; finally, an ML tree was inferred
872 under LG+C60+G+F+PMSF model (settings: -B 1000 -alrt 1000) using the guide tree
873 inferred from LG+G.
874
875

876 **Phylogenetic analyses of oligosaccharyltransferase 3/6 like (OST3/6 like) and**
877 **translocon-associated protein (TRAP) complexes subunit beta**
878
879 The OST3/6-like homologs and TRAP beta were retrieved from 966 archaeal, 1325
880 bacterial and 137 eukaryotic genomes/MAGs or largely complete transcriptomes using
881 PF04756 and PF05753 respectively, from the annotation results described above.
882 These sequences were dereplicated using cd-hit v4.7[125] (-c 0.7). Sequences of each
883 gene family were first aligned with mafft-linsi v7.453, and gappy columns were
884 removed using trimal 1.2rev59[120] with -gappyout flag. An initial ML tree was inferred
885 under LG+G using IQ-Tree v2.1.2[90] for each gene family, and sequences showing
886 potential instances of LBA artefacts were removed; remaining sequences for each
887 gene family were then realigned and trimmed as described above; finally an ML tree
888 was then inferred for each gene under LG+C60+G+F+PMSF model (settings: -B 1000
889 -alrt 1000) using guide tree inferred from LG+G.
890
891 **Phylogenetic analyses of the Snf7 family**
892 The Snf7 family homologs were identified and retrieved from 966 archaeal and 137
893 eukaryotic genomes/MAGs based on the presence of the domain IPR005024.
894 Eukaryotic sequences were dereplicated using cd-hit v4.7[120] (-c 0.7). We created two
895 datasets for this family, i.e. a dataset with only archaeal sequences and another with
896 archaeal and eukaryotic sequences. Sequences of each dataset were first aligned with
897 mafft-linsi v7.453[88], and poorly aligned sites were removed using BMGE v1.1.2[89]
898 (settings: -m BLOSUM30 -h 0.55). The inference of phylogenetic trees was conducted
899 as follows: an initial ML tree was inferred under LG+G using IQ-Tree v 2.1.2 [90], and
900 sequences showing potential instances of LBA artefacts were removed; remaining
901 sequences were then realigned and trimmed as described above; finally, an ML tree
902 was then inferred funder the LG+C60+G+F+PMSF model (settings: -B 1000 -alrt 1000)
903 using a guide tree inferred from LG+G. We also removed the 20% most
904 heterogeneous sites from the alignment with both archaeal and eukaryotic sequences
905 included using Alignment_pruner.pl
906 (https://github.com/novigit/davinciCode/blob/master/perl) and inferred an ML tree
907 using the same approach.
908
909

910 **Phylogenetic analyses of CDP-archaeol synthase (CarS), archaetidylinositol**
911 **phosphate synthase (PgsA) and phosphatidylserine synthase (PssA)**
912 Sequences identified as CarS (K19664), PgsA (K17884), and PssA (K17103) were
913 retrieved from 966 archaeal, 1325 bacterial and 137 eukaryotic genomes/MAGs or
914 largely complete transcriptomes, respectively, from the annotation results described
915 above. Sequences of each gene family were first aligned with mafft-linsi v7.453[88], and
916 gappy columns were removed using trimal 1.2rev59[120] with the -gappyout flag. The
917 inference of phylogenetic tree for individual gene families was conducted as follows:
918 an initial ML tree was inferred under LG+G using IQ-Tree v2.1.2[90] for each gene family,
919 and sequences showing potential instance of LBA artefacts were removed; remaining
920 sequences for each gene were then realigned and trimmed as described above; finally
921 an ML tree was then inferred for each gene under LG+C60+G+F+PMSF model
922 (settings: -B 1000 -alrt 1000) using guide tree inferred from LG+G.
923
924 **Phylogenetic analyses of [NiFe] group hydrogenase, large subunit**
925 Backbone sequences for phylogenetic analyses of [NiFe] group 3 and 4 hydrogenases
926 were obtained from a previous study[14]. Njordarchaeia homologs were added to the
927 backbone sequences and aligned with mafft-linsi v7.453. Poorly aligned sites were
928 removed using BMGE v1.1.2[89](settings: -m BLOSUM30 -h 0.55). A phylogenetic tree
929 was inferred under the LG+C60+G+F+PMSF model guide trees inferred from the
930 LG+G+F model using IQ-Tree v2.1.2[90] (settings: -B 1000 -alrt 1000). Phylogenetic tree
931 was visualised using ggtree 3.2.0[126].
932
933 **Network inferences**
934 The input dataset for network prediction was obtained from OTU community profiles
935 derived from 248,559 NCBI metagenomes[127] from the 'Sandpiper' website
936 (https://sandpiper.qut.edu.au), which were annotated with the GTDB R214
937 taxonomy[128]. First, we used the known association between *Huberarchaeum* and
938 *Altiarchaeum* to confirm that our network association approach is able to detect
939 symbiotic interactions. We selected all samples that contain taxa annotated as
940 *Huberarchaeum* (67 samples), respectively. We transformed the coverage of the taxa
941 in these samples to counts by multiplying with 100 and formatted the data into a table
942 listing files in rows and all OTUs in columns. To reduce the scarcity of the dataset, we
943 filtered for OTUs present in 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of the

944 samples. To estimate the minimum number of samples required for robust, true
945 positive prediction of *Huberarchaeum-Altiarchaeum* associations, we drew random
946 samples and applied two different filter criteria. Numbers of 10, 20, 30, 40, 50, and 60
947 samples were drawn; the random data sets were filtered for OTUs detected in at least
948 50% and 80% of all random samples, respectively. For each condition, ten networks
949 were inferred. Statistics were run across all ten networks within one condition. These
950 networks were inferred using SpiecEasi v1.1.3[129]. Across all networks, we checked for
951 associations between *Huberarchaeum* and *Altiarchaeum*. In all networks inferred (>30
952 samples), the expected association was robustly detected. However, below a number
953 of 30 samples, the association was detected only in 50% of the samples. This clearly
954 indicates a minimum number of > 30 samples (equaling 40% of samples) that is
955 necessary for reliable inference of association.
956 We also attempted to confirm the known association between *Nanoarchaeum* and its
957 host *Ignicoccus* (order Sulfolobales). Our dataset contained several OTUs, assigned
958 to the genus *Nanoarchaeum* (otu_658, otu_3077) and the genus *Ignicoccus*
959 (otu_3007, otu_678, otu_3062, otu_3063, otu_3922, otu_4367). However, some of the
960 taxa were present in only very few samples (**Supplementary Figure 38,39)**, and
961 subsequently in the inferred network, no association between any of these taxa was
962 predicted. Interestingly, some OTUs showed strong, significant negative correlations
963 (**Supplementary Figure 38,39**), suggesting a high specificity of the symbiotic
964 association on OTU level, although larger datasets are needed to confirm this
965 assumption.

966 Next, we selected all samples that contain taxa annotated as "*Panguiarchaeaceae"* in
967 GTDB R220[38,128], a taxonomic rank that is equivalent to the order Njordarchaeales
968 proposed in this manuscript (**Figure 1BC**), resulting in a subset of 120 samples. To
969 reduce the scarcity of the dataset, we filtered the dataset before network inference.
970 We used 5 different occupancy thresholds based on shoulder theory and commonly
971 applied criteria to prepare 5 different datasets from which separate networks were
972 inferred. OTUs had to be present in (1) at least 24 samples as well as in at least (2)
973 40%, (3) 50%, (4) 60%, and (5) 70% of the entire dataset to be included in the network.
974 Across all networks, we checked for associations involving OTUs of our target lineage
975 in a comparative manner. To exclude random associations involving our target
976 lineage, we prepared random networks and selected only those associations that were

977 unlikely to be found randomly. For each filter criterion, 999 random networks were
978 drawn based on the original network inferred from the filtered data. Then we compared
979 the taxonomic distribution of OTUs associated with OTUs annotated as
980 Njordarchaeales (equivalent to "*Panguiarchaeaceae*" in GTDB R220) between the
981 random networks and the original network. Association frequencies were compared
982 based on phylum, class, order, and family level. Next, we explored associations at the
983 genus level. A single OTU within the genus *Panguiarchaeum* was detected in higher
984 numbers across samples, that is OTU_804 (43 samples), for which we investigated
985 association partners. All analyses were performed in R v4.4.1 [130].
986

987 **Data availability**

988 All genomic data/transcriptomic data analysed in this study are available at NCBI
989 (Supplementary Data 15, 17) and are deposited in the data repository
990 (https://doi.org/10.5281/zenodo.14777939). Data generated in this study including
991 single gene and concatenated phylogenies (i.e., sequence files, alignments, and
992 treefiles) have also been deposited in our data repository at Zenodo
993 (https://doi.org/10.5281/zenodo.14777939) under the following license CC BY 4.0.

994

995 **Code availability**

996 Workflows for annotations, phylogenies, reconciliation, and custom scripts
997 (bash/python) to analyse and parse annotation data, generate single gene trees and
998 perform ancestral gene content reconstruction are deposited in the repository
999 (https://doi.org/10.5281/zenodo.14777939).

1000

1001

1002

1003

1004

1005

1006

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 947317, ASymbEL to A.S. and grant agreement No. 714774, GENECLOCKS to G.J.S.). Furthermore, this work was supported by the Gordon and Betty Moore Foundation (GBMF9741 to T.A.W., A.S., and G.J.S.) and the Simons Foundation with the Moore–Simons Project on the Origin of the Eukaryotic Cell, 735929LPI (https://doi.org/10.46714/735929LPI) (to A.S. and coworkers). Our research is funded by the John Templeton Foundation (63451 to L.Sz., G.J.Sz., T.A.W. and A.S.; the opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. We also want to thank Tara Mahendrarajah for valuable feedback on the manuscript.

Author contributions

W.H. T.A.W. and A.S. conceived the study; W.H. analysed and annotated genomes and generated accompanying data tables; W.H. and A.S. analysed genomic data; W.H. T.A.W. and A.S. performed and analyzed phylogenetic analyses, C.R. and M.P. performed co-proportionality analysis; W.H., Z.H., T.W., C.R., M.P., G.J.S, L.S, T.E., and A.S. interpreted data; W.H., T.W., and A.S wrote and all authors edited and approved the manuscript.

Competing interests

The authors declare no competing interests.

1041 **Reference**:

1042 1. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and
1043 eukaryotes. *Nature* **521**, 173–179 (2015).
1044 2. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of
1045 eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
1046 3. Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with
1047 eukaryotes. *Nature* **593**, 553–557 (2021).
1048 4. Eme, L. *et al.* Inference and reconstruction of the heimdallarchaeial ancestry of
1049 eukaryotes. *Nature* **618**, 992–999 (2023).
1050 5. Wu, F. *et al.* Unique mobile elements and scalable gene flow at the prokaryote-
1051 eukaryote boundary revealed by circularized Asgard archaea genomes. *Nat*
1052 *Microbiol* **7**, 200–212 (2022).
1053 6. Valentin-Alvarado, L. E. *et al.* Asgard archaea modulate potential
1054 methanogenesis substrates in wetland soil. *Nat Commun* **15**, 6384 (2024).
1055 7. Vosseberg, J. *et al.* The emerging view on the origin and early evolution of
1056 eukaryotic cells. *Nature* **633**, 295–305 (2024).
1057 8. Akıl, C. *et al.* Insights into the evolution of regulated actin dynamics via
1058 characterization of primitive gelsolin/cofilin proteins from Asgard archaea. *Proc.*
1059 *Natl. Acad. Sci. U. S. A.* **117**, 19904–19913 (2020).
1060 9. Hatano, T. *et al.* Asgard archaea shed light on the evolutionary origins of the
1061 eukaryotic ubiquitin-ESCRT machinery. *Nat. Commun.* **13**, 3398 (2022).
1062 10. Hurtig, F. *et al.* The patterned assembly and stepwise Vps4-mediated
1063 disassembly of composite ESCRT-III polymers drives archaeal cell division. *Sci*
1064 *Adv* **9**, eade5224 (2023).
1065 11. Akıl, C. & Robinson, R. C. Genomes of Asgard archaea encode profilins that
1066 regulate actin. *Nature* **562**, 439–443 (2018).
1067 12. Lu, Z. *et al.* Coevolution of Eukaryote-like Vps4 and ESCRT-III Subunits in the
1068 Asgard Archaea. *mBio* **11**, (2020).
1069 13. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface.
1070 *Nature* **577**, 519–525 (2020).
1071 14. Spang, A. *et al.* Proposal of the reverse flow model for the origin of the
1072 eukaryotic cell based on comparative analyses of Asgard archaeal metabolism.
1073 *Nat Microbiol* **4**, 1138–1148 (2019).

1074 15. Martin, W. & Müller, M. The hydrogen hypothesis for the first eukaryote. *Nature*
1075 **392**, 37–41 (1998).
1076 16. Moreira, D. & Lopez-Garcia, P. Symbiosis between methanogenic archaea and
1077 delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J*
1078 *Mol Evol* **47**, 517–530 (1998).
1079 17. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–
1080 934 (2010).
1081 18. Poole, A. M. & Gribaldo, S. Eukaryotic origins: How and when was the
1082 mitochondrion acquired? *Cold Spring Harb Perspect Biol* **6**, a015990 (2014).
1083 19. Koonin, E. V. & Yutin, N. The dispersed archaeal eukaryome and the complex
1084 archaeal ancestor of eukaryotes. *Cold Spring Harb Perspect Biol* **6**, a016188
1085 (2014).
1086 20. Guy, L., Saw, J. H. & Ettema, T. J. G. The archaeal legacy of eukaryotes: a
1087 phylogenomic perspective. *Cold Spring Harb Perspect Biol* **6**, a016022 (2014).
1088 21. López-García, P. & Moreira, D. Open Questions on the Origin of Eukaryotes.
1089 *Trends Ecol Evol* **30**, 697–708 (2015).
1090 22. Ettema, T. J. G. Evolution: Mitochondria in the second act. *Nature* **531**, 39–40
1091 (2016).
1092 23. Roger, A. J., Muñoz-Gómez, S. A. & Kamikawa, R. The Origin and
1093 Diversification of Mitochondria. *Curr Biol* **27**, R1177–R1192 (2017).
1094 24. Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J. & Embley, T. M.
1095 Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol*
1096 *Evol* **4**, 138–147 (2020).
1097 25. Elkins, J. G. *et al.* A korarchaeal genome reveals insights into the evolution of
1098 the Archaea. *Proc Natl Acad Sci U S A* **105**, 8102–8107 (2008).
1099 26. Tahon, G. *et al.* Phylogenomics and ancestral reconstruction of Korarchaeota
1100 reveals genomic adaptation to habitat switching. *bioRxiv* (2023)
1101 doi:10.1101/2023.09.28.559970.
1102 27. Guy, L. & Ettema, T. J. G. The archaeal 'TACK' superphylum and the origin of
1103 eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).
1104 28. Baker, B. J. *et al.* Diversity, ecology and evolution of Archaea. *Nat Microbiol* **5**,
1105 887–900 (2020).
1106 29. Dombrowski, N. *et al.* Undinarchaeota illuminate DPANN phylogeny and the
1107 impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).

1108 30. Xie, R. *et al.* Expanding Asgard members in the domain of Archaea sheds new
1109 light on the origin of eukaryotes. *Sci. China Life Sci.* **65**, 818–829 (2022).
1110 31. Liu, Y. & Li, M. The unstable evolutionary position of Korarchaeota and its
1111 relationship with other TACK and Asgard archaea. *mLife* **1**, 218–222 (2022).
1112 32. Qu, Y.-N. *et al.* Panguiarchaeum symbiosum, a potential hyperthermophilic
1113 symbiont in the TACK superphylum. *Cell Rep.* **42**, 112158 (2023).
1114 33. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark
1115 matter. *Nature* **499**, 431–437 (2013).
1116 34. Castelle, C. J. *et al.* Biosynthetic capacity, metabolic variety and unusual biology
1117 in the CPR and DPANN radiations. *Nat. Rev. Microbiol.* **16**, 629–645 (2018).
1118 35. Tamarit, D. *et al.* Description of Asgardarchaeum abyssi gen. nov. spec. nov., a
1119 novel species within the class Asgardarchaeia and phylum Asgardarchaeota in
1120 accordance with the SeqCode. *Syst Appl Microbiol* **47**, 126525 (2024).
1121 36. Imachi, H. *et al.* gen. nov., sp. nov., an anaerobic, obligately syntrophic
1122 archaeon, the first isolate of the lineage 'Asgard' archaea, and proposal of the
1123 new archaeal phylum phyl. nov. and kingdom regn. nov. *Int J Syst Evol*
1124 *Microbiol* **74**, (2024).
1125 37. Moody, E. R. R. *et al.* An estimate of the deepest branches of the tree of life
1126 from ancient vertically evolving genes. *Elife* **11**, (2022).
1127 38. Rinke, C. *et al.* A standardized archaeal taxonomy for the Genome Taxonomy
1128 Database. *Nat Microbiol* **6**, 946–959 (2021).
1129 39. Lartillot, N. & Philippe, H. A Bayesian mixture model for across-site
1130 heterogeneities in the amino-acid replacement process. *Mol Biol Evol* **21**, 1095–
1131 1109 (2004).
1132 40. Williams, T. A. *et al.* Inferring the Deep Past from Molecular Data. *Genome Biol*
1133 *Evol* **13**, (2021).
1134 41. Kapli, P., Flouri, T. & Telford, M. J. Systematic errors in phylogenetic trees.
1135 *Current Biology* **31**, R59–R64 (2021).
1136 42. Zeldovich, K. B., Berezovsky, I. N. & Shakhnovich, E. I. Protein and DNA
1137 sequence determinants of thermophilic adaptation. *PLoS Comput Biol* **3**, e5
1138 (2007).
1139 43. Szilágyi, A. & Závodszky, P. Structural differences between mesophilic,
1140 moderately thermophilic and extremely thermophilic protein subunits: results of
1141 a comprehensive survey. *Structure* **8**, 493–504 (2000).

1142 44. Cambillau, C. & Claverie, J. M. Structural and genomic correlates of
1143 hyperthermostability. *J Biol Chem* **275**, 32383–32386 (2000).
1144 45. Muñoz-Gómez, S. A. *et al.* Site-and-branch-heterogeneous analyses of an
1145 expanded dataset favour mitochondria as sister to known Alphaproteobacteria.
1146 *Nat Ecol Evol* **6**, 253–262 (2022).
1147 46. Glasemacher, J., Bock, A. K., Schmid, R. & Schønheit, P. Purification and
1148 properties of acetyl-CoA synthetase (ADP-forming), an archaeal enzyme of
1149 acetate formation and ATP synthesis, from the hyperthermophile Pyrococcus
1150 furiosus. *Eur J Biochem* **244**, 561–567 (1997).
1151 47. Yu, H. *et al.* Structure of an Ancient Respiratory System. *Cell* **173**, 1636–
1152 1649.e16 (2018).
1153 48. Brown, A. M., Hoopes, S. L., White, R. H. & Sarisky, C. A. Purine biosynthesis in
1154 archaea: variations on a theme. *Biol Direct* **6**, 63 (2011).
1155 49. Nishimura, Y. & Eguchi, T. Biosynthesis of archaeal membrane lipids:
1156 digeranylgeranylglycerophospholipid reductase of the thermoacidophilic
1157 archaeon Thermoplasma acidophilum. *J. Biochem.* **139**, 1073–1081 (2006).
1158 50. Williams, T. A. *et al.* Phylogenetic reconciliation: making the most of genomes to
1159 understand microbial ecology and evolution. *ISME J* **18**, (2024).
1160 51. Baker, B. A. *et al.* Expanded phylogeny of extremely halophilic archaea shows
1161 multiple independent adaptations to hypersaline environments. *Nature*
1162 *Microbiology* **9**, 964–975 (2024).
1163 52. Huang, W.-C. *et al.* Comparative genomic analysis reveals metabolic flexibility
1164 of Woesearchaeota. *Nature Communications* **12**, 1–14 (2021).
1165 53. Williams, T. A. *et al.* Integrative modeling of gene and genome evolution roots
1166 the archaeal tree of life. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4602–E4611
1167 (2017).
1168 54. Martijn, J. *et al.* Hikarchaeia demonstrate an intermediate stage in the
1169 methanogen-to-halophile transition. *Nat Commun* **11**, 5490 (2020).
1170 55. Forterre, P. A hot story from comparative genomics: reverse gyrase is the only
1171 hyperthermophile-specific protein. *Trends Genet* **18**, 236–237 (2002).
1172 56. Lu, Z. *et al.* Evolution of optimal growth temperature in Asgard archaea inferred
1173 from the temperature dependence of GDP binding to EF-1A. *Nature*
1174 *Communications* **15**, 1–7 (2024).
1175 57. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth temperature and

genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biol Evol* **5**, 966–977 (2013).

58. Pierpont, C. L., Baroch, J. J., Church, M. J. & Miller, S. R. Idiosyncratic genome evolution of the thermophilic cyanobacterium Synechococcus at the limits of phototrophy. *ISME J* **18**, (2024).
59. McCutcheon, J. P. & Moran, N. A. Extreme genome reduction in symbiotic bacteria. *Nature Reviews Microbiology* **10**, 13–26 (2011).
60. Huber, H. *et al.* A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* **417**, 63–67 (2002).
61. Schwank, K. *et al.* An archaeal symbiont-host association from the deep terrestrial subsurface. *ISME J* **13**, 2135–2139 (2019).
62. Munson-McGee, J. H. *et al.* Nanoarchaeota, Their Sulfolobales Host, and Nanoarchaeota Virus Distribution across Yellowstone National Park Hot Springs. *Appl Environ Microbiol* **81**, 7860–7868 (2015).
63. Sakai, H. D. *et al.* Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. *Proc Natl Acad Sci U S A* **119**, (2022).
64. Kato, S. *et al.* gen. nov., sp. nov., a thermoacidophilic, obligate ectosymbiotic archaeon, and proposal of fam. nov., ord. nov. and class. nov. *Int J Syst Evol Microbiol* **72**, (2022).
65. St John, E. *et al.* A new symbiotic nanoarchaeote (Candidatus Nanoclepta minutus) and its host (Zestosphaera tikiterensis gen. nov., sp. nov.) from a New Zealand hot spring. *Syst Appl Microbiol* **42**, 94–106 (2019).
66. Wurch, L. *et al.* Genomics-informed isolation and characterization of a symbiotic Nanoarchaeota system from a terrestrial geothermal environment. *Nat Commun* **7**, 12115 (2016).
67. Podar, M. *et al.* Insights into archaeal evolution and symbiosis from the genomes of a nanoarchaeon and its inferred crenarchaeal host from Obsidian Pool, Yellowstone National Park. *Biol Direct* **8**, 9 (2013).
68. Schut, G. J., Boyd, E. S., Peters, J. W. & Adams, M. W. W. The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. *FEMS Microbiol Rev* **37**, 182–203 (2013).
69. Topçuoğlu, B. D. *et al.* Hydrogen limitation and syntrophic growth among natural assemblages of thermophilic methanogens at deep-sea hydrothermal vents.

1210 *Front. Microbiol.* **7**, 1240 (2016).
1211 70. Ver Eecke, H. C. *et al.* Hydrogen-limited growth of hyperthermophilic
1212 methanogens at deep-sea hydrothermal vents. *Proc Natl Acad Sci U S A* **109**,
1213 13674–13679 (2012).
1214 71. Yu, T. *et al.* Thermophilic Hadarchaeota grow on long-chain alkanes in
1215 syntrophy with methanogens. *Nat Commun* **15**, 6560 (2024).
1216 72. Ragsdale, S. W. & Pierce, E. Acetogenesis and the Wood-Ljungdahl pathway of
1217 CO(2) fixation. *Biochim Biophys Acta* **1784**, 1873–1898 (2008).
1218 73. Schuchmann, K. & Müller, V. Autotrophy at the thermodynamic limit of life: a
1219 model for energy conservation in acetogenic bacteria. *Nat Rev Microbiol* **12**,
1220 809–821 (2014).
1221 74. Schuchmann, K. & Müller, V. Energetics and Application of Heterotrophy in
1222 Acetogenic Bacteria. *Appl Environ Microbiol* **82**, 4056–4069 (2016).
1223 75. Huber, H. *et al.* Ignicoccus gen. nov., a novel genus of hyperthermophilic,
1224 chemolithoautotrophic Archaea, represented by two new species, Ignicoccus
1225 islandicus sp nov and Ignicoccus pacificus sp nov. and Ignicoccus pacificus sp.
1226 nov. *Int J Syst Evol Microbiol* **50 Pt 6**, 2093–2100 (2000).
1227 76. Sakai, H. D., Nakamura, K. & Kurosawa, N. sp. nov., a facultatively anaerobic
1228 hyperthermophilic archaeon isolated from the Unzen hot spring in Japan. *Int J*
1229 *Syst Evol Microbiol* **72**, (2022).
1230 77. Leung, P. M. *et al.* Trace gas oxidation sustains energy needs of a thermophilic
1231 archaeon at suboptimal temperatures. *Nat Commun* **15**, 3219 (2024).
1232 78. López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of
1233 eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).
1234 79. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Deep
1235 mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101–
1236 105 (2018).
1237 80. Molina, C. S., Williams, T. A., Snel, B. & Spang, A. Chimeric origins and
1238 dynamic evolution of central carbon metabolism in eukaryotes. *bioRxiv* (2024)
1239 doi:10.1101/2024.05.29.596406.
1240 81. Freibert, S.-A. *et al.* Evolutionary conservation and in vitro reconstitution of
1241 microsporidian iron–sulfur cluster biosynthesis. *Nature Communications* **8**, 1–12
1242 (2017).
1243 82. Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with

1244 ancient duplications. *Nat Ecol Evol* **5**, 92–100 (2021).
1245 83. Pittis, A. A. & Gabaldón, T. Late acquisition of mitochondria by a host with
1246 chimaeric prokaryotic ancestry. *Nature* **531**, 101–104 (2016).
1247 84. Eme, L., Spang, A., Lombard, J., Stairs, C. W. & Ettema, T. J. G. Archaea and
1248 the origin of eukaryotes. *Nat Rev Microbiol* **15**, 711–723 (2017).
1249 85. Tria, F. D. K. *et al.* Gene Duplications Trace Mitochondria to the Onset of
1250 Eukaryote Complexity. *Genome Biol Evol* **13**, (2021).
1251 86. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive
1252 sequence similarity searching. *Nucleic Acids Res* **39**, W29–37 (2011).
1253 87. Werner, F. Structure and function of archaeal RNA polymerases. *Mol. Microbiol.*
1254 **65**, 1395–1404 (2007).
1255 88. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software
1256 version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–
1257 780 (2013).
1258 89. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy):
1259 a new software for selection of phylogenetic informative regions from multiple
1260 sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
1261 90. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for
1262 Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 1530–1534
1263 (2020).
1264 91. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile
1265 HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
1266 92. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **32**,
1267 D138–41 (2004).
1268 93. Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model
1269 organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281
1270 (2021).
1271 94. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous
1272 Genes (arCOGs): An Update and Application for Analysis of Shared Features
1273 between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**,
1274 818–840 (2015).
1275 95. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin,
1276 L. S. ModelFinder: fast model selection for accurate phylogenetic estimates.
1277 *Nat. Methods* **14**, 587–589 (2017).

1278 96. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity
1279 with Posterior Mean Site Frequency Profiles Accelerates Accurate
1280 Phylogenomic Estimation. *Syst Biol* **67**, 216–235 (2018).
1281 97. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes
1282 substantially expands the tree of life. *Nat Microbiol* **2**, 1533–1542 (2017).
1283 98. Szánthó, L. L., Lartillot, N., Szöllősi, G. J. & Schrempf, D. Compositionally
1284 Constrained Sites Drive Long-Branch Attraction. *Syst. Biol.* **72**, 767–780 (2023).
1285 99. Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: phylogenetic
1286 reconstruction with infinite mixtures of profiles in a parallel environment. *Syst.*
1287 *Biol.* **62**, 611–615 (2013).
1288 100. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior
1289 Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–
1290 904 (2018).
1291 101. Schrempf, D., Lartillot, N. & Szöllősi, G. Scalable Empirical Mixture Models That
1292 Account for Across-Site Compositional Heterogeneity. *Mol. Biol. Evol.* **37**, 3616–
1293 3631 (2020).
1294 102. Susko, E. & Roger, A. J. On Reduced Amino Acid Alphabets for Phylogenetic
1295 Inference. *Mol. Biol. Evol.* **24**, 2139–2150 (2007).
1296 103. Kosiol, C., Goldman, N. & Buttimore, N. H. A new criterion and method for
1297 amino acid classification. *J. Theor. Biol.* **228**, 97–106 (2004).
1298 104. Shimodaira, H. Multiple comparisons of log-likelihoods and combining
1299 nonnested models with applications to phylogenetic tree selection. *Commun.*
1300 *Stat. Theory Methods* **30**, 1751–1772 (2001).
1301 105. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome
1302 phylogeny substantially revises the tree of life. *Nat Biotechnol* **36**, 996–1004
1303 (2018).
1304 106. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv [cs.LG]*
1305 (2012) doi:10.48550/ARXIV.1201.0490.
1306 107. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**,
1307 3021 (2021).
1308 108. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**,
1309 2068–2069 (2014).
1310 109. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B.
1311 The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*

1312 **42**, D490–5 (2014).
1313 110. Saier, M. H. *et al.* The Transporter Classification Database (TCDB): 2021
1314 update. *Nucleic Acids Res.* **49**, D461–D467 (2021).
1315 111. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for
1316 hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).
1317 112. Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their
1318 substrates and inhibitors in 2017 and a comparison with peptidases in the
1319 PANTHER database. *Nucleic Acids Res.* **46**, D624–D632 (2018).
1320 113. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
1321 *Bioinformatics* **30**, 1236–1240 (2014).
1322 114. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of
1323 protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
1324 115. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
1325 DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
1326 116. The pandas development team. *Pandas-Dev/pandas: Pandas*. (Zenodo, 2024).
1327 doi:10.5281/ZENODO.3509134.
1328 117. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual: (Python*
1329 *Documentation Manual Part 2)*. (CreateSpace, 2009).
1330 118. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive
1331 sequence similarity searching. *Nucleic Acids Res* **39**, W29–37 (2011).
1332 119. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing
1333 genomic features. *Bioinformatics* **26**, 841–842 (2010).
1334 120. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for
1335 automated alignment trimming in large-scale phylogenetic analyses.
1336 *Bioinformatics* **25**, 1972–1973 (2009).
1337 121. Davín, A. A. *et al.* An evolutionary timescale for Bacteria calibrated using the
1338 Great Oxidation Event. *bioRxiv* (2023) doi:10.1101/2023.08.08.552427.
1339 122. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: a rapid,
1340 scalable and accurate tool for assessing microbial genome quality using
1341 machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
1342 123. Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-
1343 scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos*
1344 *Trans R Soc Lond B Biol Sci* **370**, 20140335 (2015).
1345 124. Coleman, G. A. *et al.* A rooted phylogeny resolves early bacterial evolution.

Science **372**, (2021).
125. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
126. Xu, S. *et al.* : A serialized data object for visualization of a phylogenetic tree and annotation data. *Imeta* **1**, e56 (2022).
127. Woodcroft, B. J. *et al.* SingleM and Sandpiper: Robust microbial taxonomic profiles from metagenomic data. *bioRxiv* 2024.01.30.578060 (2024) doi:10.1101/2024.01.30.578060.
128. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* **50**, D785–D794 (2022).
129. Kurtz, Z. D. *et al.* Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology* **11**, e1004226 (2015).
130. R Core Team R. R: A language and environment for statistical computing. *Paperpile* https://paperpile.com/shared/sVIc4WSqzTNicxf4zfapdZw (2024).

Figures and legends

Figure 1

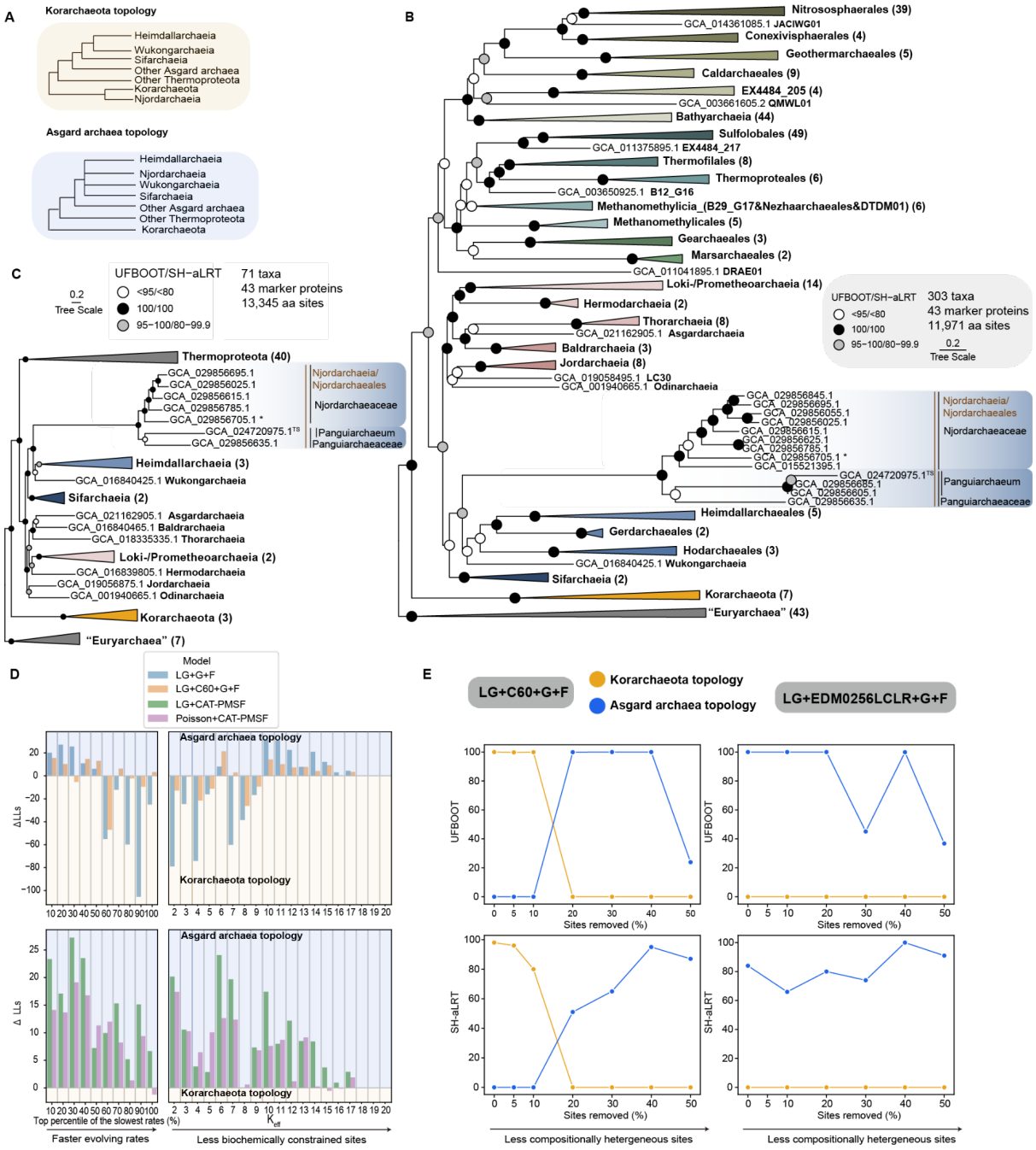
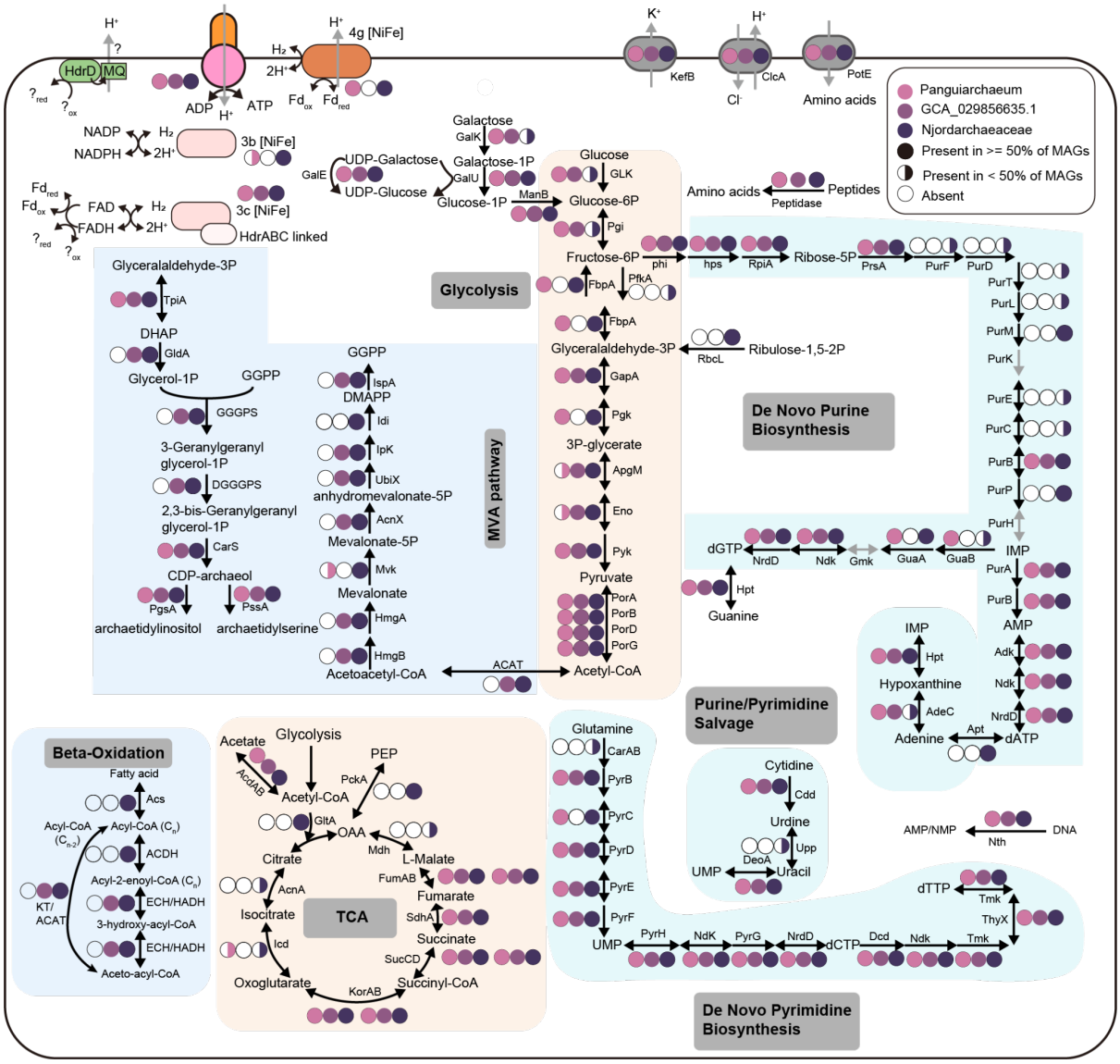


Figure 1. The phylogenetic placement of Njordarchaeia. A. Schematic of the phylogenetic hypotheses tested in this study. Korarchaeota topology: Njordarchaeia branch sister to Korarchaeota. Asgard archaea topology: Njordarchaeia branch within

Asgard archaea. **B**. Maximum-likelihood phylogenetic analysis of the concatenated 50% top-ranked marker proteins (n=43) and streamlined set2 (303 taxa) using a custom site-heterogeneous substitution model (LG+EDM0256LCLR+G+F). Scale bar: average substitutions per site. The number of taxa in each collapsed clade is shown by the number in parenthesis next to the clade name. GCA_029856705.1* is the closest genome to the first proposed type material of *Njordarchaeum guaymasis* (See Supplementary Discussion). The GCA_024720975.1^{TS} is the proposed type material for *P.* symposium. The tree is rooted by "Euryarchaea" (including Halobacteriota, Thermoplasmatota, Hydrothermarchaeota, Methanobacteriota, Methanobacteriota_A and Methanobacteriota_B). **C**. Maximum-likelihood phylogenetic analysis of the concatenated 50% top-ranked marker proteins (n=43) and streamlined set3 (71 taxa) using LG+CAT-PMSF model. Scale bar: average substitutions per site. The number of taxa in each collapsed clade is shown by the number in parenthesis next to the clade name. **D**. Evaluation of phylogenetic signals for Asgard archaea and Korarchaeota topology under LG+G+F, LG+C60+G+F, Poisson+CAT-PMSF and LG+CAT-PMSF models. Sites underlying the signals were binned by substitution rates (based on LG+C60+G+F) and effective amino acids per site (based on LG+CAT-PMSF). **E.** Impact of progressive removal of the most compositionally biased sites ranked by chi-square score on the statistical support of Asgard archaea and Korarchaeota topology. The support values, UFBOOT and SH-aLRT, were estimated on the 50% top-ranked markers of streamlined set2 (303 taxa) based on LG+C60+G+F and LG+EDM0256LCLR+G+F model.

1413 **Figure 2**

1414

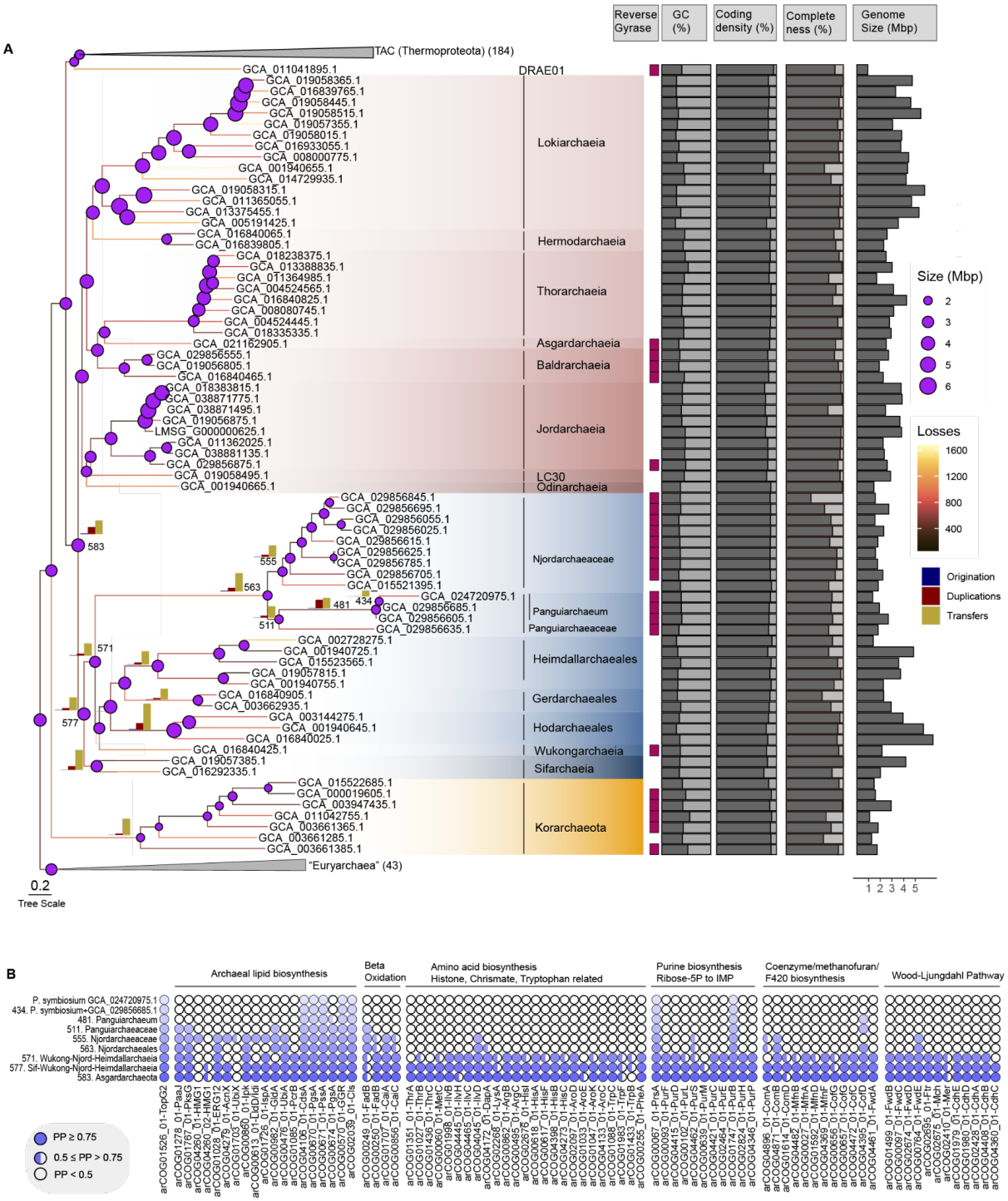


1415

1416 **Figure 2. Metabolic characteristics of Njordarchaeia.** Overall, Njordarchaeia are
1417 thermophilic fermentative heterotrophs, with core pathways of varying completeness
1418 suggesting some core metabolites must be obtained from other organisms in the
1419 environment. Full circles indicate that a gene is present in all or more than 50% of the
1420 MAGs, while half-circles indicate that a gene is found in at least one, but less than half,
1421 of MAGs. Open circles denote genes absent from all Njordarchaeia MAGs. A detailed
1422 list of genes encoded by Njordarchaeia can be found in Supplementary Data 7,8,9.

1423

1424 **Figure 3**



1425

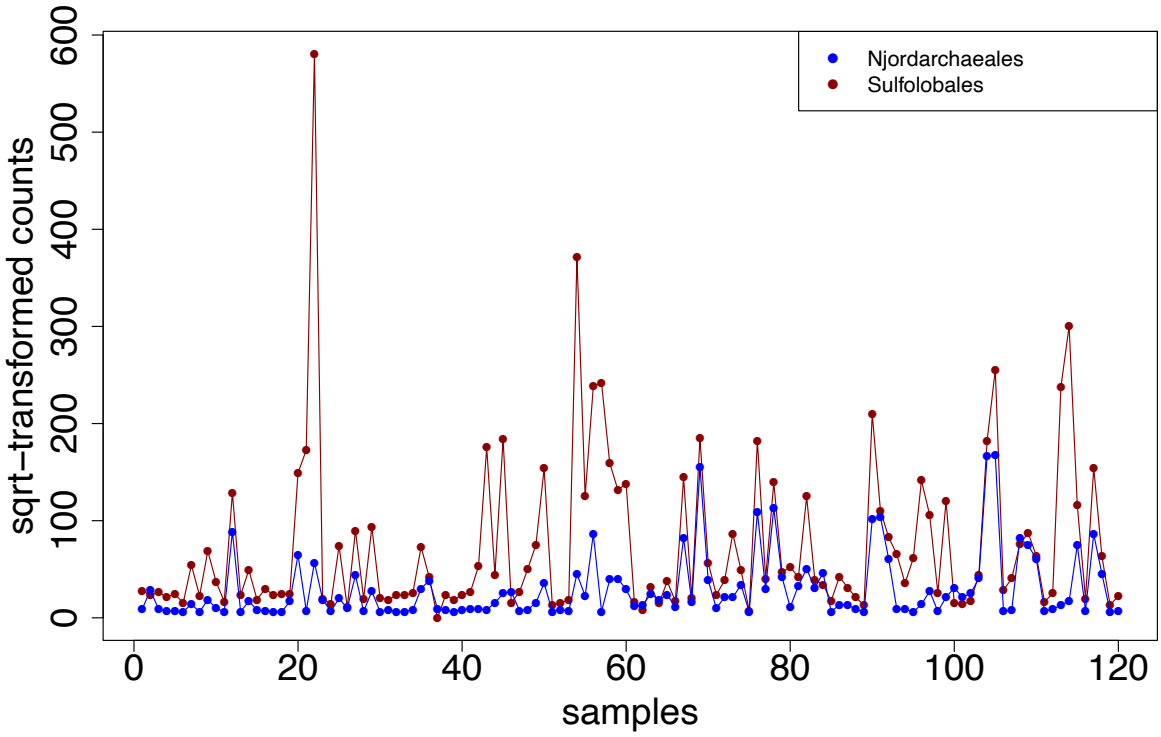
1426

1427 **Figure 3. Gene-tree-aware ancestral reconstruction of *Panguiarchaeum***
1428 **lineages.** Njordarchaeia, in particular *Panguiarchaeum*, might have experienced
1429 genome reduction from their common ancestor with a larger ancestor

Wukongarchaeia and Heimdallarchaeia (0.53-0.84 Mpb on the stem: node 571 to node 563). A. Inferred species of asgard archaeal, thermoproteotal and korarchaeotal lineages from Figure 1B. The colours on the branches represent the number of gene losses, and the bar plots on the branches of interest show the number of originations, duplications and transfers. The size of circles at nodes represents the inferred ancestral genome size. The number of taxa in each collapsed clade is shown by the number in parenthesis next to the clade name. B. The presence probability of genes of interest at the ancestral nodes leading to *Panguiarchaeum*. Node numbers of ancestral nodes are shown and referred to in Panel A. Full circles indicate a presence probability (PP) ≥ 0.75, half circles show a PP< 0.75 but ≥ 0.5, while PP<0.5 is shown in open circles (See Supplementary Data 12).

1446 **Figure 4**

1447



1448

1449 **Figure 4. Abundance-based co-occurrence of Njordarchaeales and**
1450 **Sulfolobales.** For plotting, the counts of both orders Njordarchaeales (including
1451 Panguiarcheaceae) and Sulfolobales, were summed in each sample and plotted. For
1452 visualisation purposes, the counts were square root transformed.

1453

1454