Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes

Laura Eme[1,2]▲, Daniel Tamarit[1,3,4]▲†, Eva F. Caceres[1,3]▲, Courtney W. Stairs[1]††, Valerie De Anda[5], Max E. Schön[1], Kiley W. Seitz[5]†††, Nina Dombrowski[5]††††, William H. Lewis[1,3]†††††, Felix Homa[3], Jimmy H. Saw[1]††††††, Jonathan Lombard[1], Takuro Nunoura[6], Wen-Jun Li[7], Zheng-Shuang Hua[8], Lin-Xing Chen[9], Jillian F. Banfield[9,10], Emily St John[11], Anna-Louise Reysenbach[11], Matthew B. Stott[12], Andreas Schramm[13], Kasper U. Kjeldsen[13], Andreas P. Teske[14], Brett J. Baker[5,15], Thijs J. G. Ettema[1,3]*

[1]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, 75123 Uppsala, Sweden

[2]Laboratoire Écologie, Systématique, Évolution, CNRS, Université Paris-Saclay, AgroParisTech, 91190 Gif-sur-Yvette, France

[3]Laboratory of Microbiology, Wageningen University and Research, 6708 WE Wageningen, The Netherlands

[4]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, SE-75007 Uppsala, Sweden

[5]Department of Marine Science, Marine Science Institute, University of Texas Austin, Port Aransas, TX, 78373, USA

[6]Research Center for Bioscience and Nanoscience (CeBN), Japan Agency for Marine-Earth Science and Technology (JAMSTEC), 2-15 Natsushima-cho, Yokosuka, 237-0061, Japan

[7]State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources and Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, PR China

[8]Chinese Academy of Sciences Key Laboratory of Urban Pollutant Conversion, Department of Environmental Science and Engineering, University of Science and Technology of China, Hefei, 230026, PR China

[9]Department of Earth and Planetary Sciences, University of California, Berkeley, California, USA

[10]Department of Environmental Science, Policy, and Management, University of California, Berkeley, California, USA

[11]Department of Biology, Portland State University, Portland, Oregon, USA

[12]School of Biological Sciences, University of Canterbury, Christchurch, 8142, New Zealand

[13]Section for Microbiology, Department of Biology, Aarhus University, 8000 Aarhus, Denmark

[14]Department of Earth, Marine and Environmental Sciences, University of North Carolina, Chapel Hill, USA

[15]Department of Integrative Biology, University of Texas Austin, TX USA

*Correspondence to: thijs.ettema@wur.nl

▲Equal contribution

† Current address: Theoretical Biology and Bioinformatics, Department of Biology, Faculty of Science, Utrecht University, Padualaan 8, 3584CH Utrecht, The Netherlands

†† Current address: Department of Biology, Lund University, Sölvegatan 35, 223 62 Lund, Sweden

††† Current address: Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany

†††† Current address: NIOZ, Royal Netherlands Institute for Sea Research, Department of Marine Microbiology and Biogeochemistry; AB Den Burg, The Netherlands.

††††† Current address: Department of Biochemistry, University of Cambridge, Cambridge, CB2 1QW, UK

††††††: Current address: Department of Biological Sciences, The George Washington University, Washington, DC, USA

Abstract

1

2 **In the ongoing debates about eukaryogenesis, the series of evolutionary events leading to the**

3 **emergence of the eukaryotic cell from prokaryotic ancestors, members of the Asgard archaea**

4 **play a key role as the closest archaeal relatives of eukaryotes. However, the nature and**

5 **phylogenetic identity of the last common ancestor of Asgard archaea and eukaryotes remain**

6 **unresolved. Here, we analyze distinct phylogenetic marker datasets of an expanded genomic**

7 **sampling of Asgard archaea and evaluate competing evolutionary scenarios using state-of-**

8 **the-art phylogenomic approaches. We find that eukaryotes are placed, with high confidence,**

9 **as a well-nested clade within Asgard archaea, as a sister lineage to Hodarchaeales, a newly**

10   **proposed order within Heimdallarchaeia. Using sophisticated gene tree/species tree**

11   **reconciliation approaches, we show that, in analogy to the evolution of eukaryotic genomes,**

12   **genome evolution in Asgard archaea involved significantly more gene duplication and fewer**

13   **gene loss events compared to other archaea. Finally, we infer that the last common ancestor**

14   **of Asgard archaea likely was a thermophilic chemolithotroph, and that the lineage from**

15   **which eukaryotes evolved adapted to mesophilic conditions and acquired the genetic**

16   **potential to support a heterotrophic lifestyle. Our work provides key insights into the**

17   **prokaryote-to-eukaryote transition and the platform for the emergence of cellular**

18   **complexity in eukaryotic cells.**

19

20   Main

21   Understanding how complex eukaryotic cells emerged from prokaryotic ancestors represents a

22   major challenge in biology[1,2]. A main point of contention in refining eukaryogenesis scenarios

23   revolves around the exact phylogenetic relationship between Archaea and eukaryotes. The use of

24   phylogenomic approaches with improved models of sequence evolution combined with a much-

25   improved archaeal taxon sampling – progressively unveiled by metagenomics – has recently

26   yielded strong support for the "two-domain" tree of life, in which the eukaryotic clade branches

27   from within Archaea[3–8]. The discovery of the first Lokiarchaeia genome provided additional

28   evidence for the two-domain topology since this lineage was shown to represent, at the time, the

29   closest relative of eukaryotes in phylogenomic analyses[9]. Moreover, Lokiarchaeia genomes were

30   found to uniquely contain many genes encoding eukaryotic signature proteins (ESPs) –proteins

31   involved in hallmark complex processes of the eukaryotic cell–, more so than any other prokaryotic

32   lineage. The subsequent identification and analyses of several diverse relatives of Lokiarchaeia,

2

33   together forming the Asgard archaea superphylum, confirmed that Asgard archaea represented the

34   closest archaeal relatives of eukaryotes[2,9,10]. Their exact evolutionary relationship to eukaryotes,

35   however, remained unresolved: it has been unclear whether eukaryotes evolved from *within*

36   Asgard archaea, or if they represented their sister-lineage[10]. Furthermore, two studies questioned

37   this view of the tree of life altogether, suggesting that Asgard archaea represent a deep-branching

38   Euryarchaeota-related clade[11,12], and that, in accordance with the "three-domain" tree, eukaryotes

39   represent a sister group to all Archaea, although this was challenged[13,14]. A follow-up study that

40   included an expanded taxonomic sampling of Asgard archaeal genome data failed to resolve the

41   phylogenetic position of eukaryotes in the tree of life[15].

42   Here, we expand the genomic diversity of Asgard archaea by generating 63 novel Asgard

43   metagenomic-assembled genomes (MAGs) from samples from 11 locations around the world. By

44   analyzing the improved genomic sampling of Asgard archaea using state-of-the-art

45   phylogenomics, including recently developed gene tree/species tree reconciliation approaches for

46   ancestral genome content reconstruction, we firmly place eukaryotes nested within the Asgard

47   archaea. By revealing key features regarding the identity, nature and physiology of the last Asgard

48   archaea and eukaryotes common ancestor (LAECA), our results represent important, thus far

49   missing pieces of the elusive eukaryogenesis puzzle.

50

51   **Expanded Asgard archaea genomic diversity**

52   To increase the genomic diversity of Asgard archaea, we sampled aquatic sediments and

53   hydrothermal deposits from eleven geographically distinct sites (Supplementary Table 1,

54   Supplementary Figure 1). After extraction and sequencing of total environmental DNA, we

55   assembled and binned metagenomic reads into MAGs. Of these MAGs, 63 were found to belong

56   to the Asgard archaea superphylum, with estimated median completeness and redundancy of 83%

57   and 4.2%, respectively (Supplementary Table 1). To assess the genomic diversity in this dataset,

58   we reconstructed a phylogeny of ribosomal proteins encoded in a conserved 15-ribosomal protein

59   (RP15) gene cluster[16] from these MAGs, and all publicly available Asgard archaea assemblies

60   (retrieved June 29th, 2021; Figure 1). These analyses expand the genomic sampling across

61   previously described major Asgard archaea clades (i.e., Loki-, Thor-, Heimdall-, Odin-, Hel-,

62   Hermod-, Sif-, Jord- and Baldrarchaeia[9,10,15,17,18]) and recover a previously undescribed clade of

63   high taxonomic rank (Candidatus Asgardarchaeia; see Ext. Data Fig. 1 and Supplementary

64   Information for proposed uniformization of Asgard archaea taxonomic classification that will be

65   adhered to throughout the present manuscript). We observed that the median estimated Asgard

66   archaeal genome size (3.8 Mega basepairs (Mbp)) is considerably larger than those of

67   representative genomes from TACK archaea and Euryarchaeota (median=1.8 Mbp for both) and

68   DPANN archaea (median=1.2 Mbp) (Supplementary Table 1). Among Asgard archaea,

69   Odinarchaeia display the smallest genomes (median=1.4 Mb), while Loki- and Helarchaeales

70   contain the largest (median=4.3 Mbp for both). Unlike other major Asgard archaeal clades,

71   Heimdallarchaeia possess a wide range of genome sizes, spanning from 1.6 to 7.4 Mbp

72   (median=3.5 Mbp). Indeed, this large class contains five clades with diverse features. These

73   include Njordarchaeales (median genome size=2.4 Mbp) followed by Kariarchaeaceae (median

74   genome size=2.7 Mbp), Gerdarchaeales (median genome size=3.4 Mbp), Heimdallarchaeaceae

75   (median genome size=3.7 Mbp), and finally Hodarchaeales (median genome size=5.1 Mbp). The

76   smallest heimdallarchaeial genome corresponds to the only Asgard archaeal MAG recovered from

77   a marine surface water metagenome (Heimdallarchaeota archaeon RS678)[19], in agreement with

78   reduced genome sizes typically observed among prokaryotic plankton of the euphotic zone[20] .

4

79

80 **Identification of phylogenetic conflict**

81 Inferring deep evolutionary relationships in the tree of life is considered one of the hardest

82 problems in phylogenetics. To interrogate the evolutionary relationships within the present set of

83 Asgard archaeal phyla, and between Asgard archaea and eukaryotes, we performed an exhaustive

84 range of sophisticated phylogenomic analyses. We analyzed a preexisting marker dataset

85 comprising 56 concatenated ribosomal protein sequences (RP56)[9,10] for a phylogenetically diverse

86 set of 331 archaeal (175 Asgard archaea, 41 DPANN, 43 Euryarchaeota, and 72 TACK archaea

87 representatives), and 14 eukaryotic taxa (see Supplementary Table 2). Of note, the inclusion of an

88 expanded diversity of 12 new Korarchaeota MAGs among these TACK archaea considerably

89 affected phylogenomic analyses (see below). Initial maximum-likelihood (ML) phylogenetic

90 inference based on this RP56 dataset confirmed the existence of 12 major Asgard archaeal clades

91 of high taxonomic rank (Supplementary Figure 2). These include the previously described Loki-,

92 Odin-, Heimdall-, Thor-[9,10], Helarchaeia[10], for which we here present 36 new genomes, and the

93 recently proposed Sif-[18], Hermod-[17], Jord-[21], Wukong-[15] and Baldrarchaeia[15], for most of which

94 we also identified new near-complete MAGs. Finally, we identified 15 MAGs representing the

95 recently described Njordarchaeales[22] (which we show below is a divergent candidate order within

96 Heimdallarchaeia), and a single MAG representing a new candidate class, Asgardarchaeia (which

97 will be discussed in a separate manuscript; Tamarit et al, in prep) (Figure 1). Importantly, careful

98 inspection of the obtained RP56 tree uncovered a potential artefact: Njordarchaeales, considered

99 *bona fide* Asgard archaea based on the presence of many encoded 'typical' Asgard-like ESPs[10],

100 were found to branch outside of the Asgard archaea, at the base of the TACK superphylum and as

101 a sister lineage to Korarchaeota in the RP56 tree. In addition, eukaryotes were found to branch at

5

102   the base of the clade formed by Korarchaeota and Njordarchaeales, although with weak support.

103   Hereafter, we focused on disentangling the historically correct phylogenetic signal from noise and

104   artefacts.

105

106   **Alternative phylogenomic markers**

107   Despite often being used in phylogenomic analyses, ribosomal proteins have been suggested to

108   contribute to phylogenetic artefacts due to inherent compositional sequence biases[23,24].

109   Additionally, considering the inconsistency of the obtained placement of eukaryotes compared to

110   previous analyses, the incoherent placement of Njordarchaeales, and the presence of long branches

111   at the base of both of these clades in the RP56 tree, we sought to use an alternative phylogenetic

112   marker set to obtain a stable Asgard archaeal species tree, and to further investigate the

113   phylogenetic position of eukaryotes. We constructed an independent 'new marker' dataset

114   comprising 57 proteins of archaeal origin in eukaryotes (NM57 dataset; see Methods). The NM57

115   proteins are mostly involved in diverse informational, metabolic, and cellular processes, but do

116   not include ribosomal proteins (Supplementary Table 2). Besides being longer, and hence

117   putatively more phylogenetically informative compared to the RP56 markers, the broader

118   functional distribution of NM57 markers is less likely to cause phylogenetic reconstruction

119   artefacts induced by strong co-evolution between proteins – something that is to be expected for

120   functionally and structurally cohesive ribosomal proteins[25]. Indeed, in case co-evolving protein

121   sequences are compositionally biased, and hence violate evolutionary model assumptions of fixed

122   composition over species, their concatenation is expected to strengthen the artefactual, non-

123   phylogenetic signal and the statistical support for incorrect relationships[26]. We thus decided to

124   independently evaluate the concatenated NM57 and RP56 marker datasets for downstream

6

125   phylogenomic analyses. We observed that ML phylogenomic analyses of the NM57 dataset did

126   not only recover Njordarchaeales as *bona fide* Asgard archaea, they were also placed as the closest

127   relatives of eukaryotes (BS=98%; Supplementary Figure 3), as was proposed in a recent analysis[22].

128   To investigate the underlying causes for the contradicting results between the NM57 and RP56

129   datasets, we first assessed the effect of taxon sampling on phylogenetic reconstructions by

130   removing eukaryotic and/or DPANN and/or Korarchaeota sequences from the alignments, for two

131   main reasons: (1) eukaryotes and DPANN archaea represent long-branching clades potentially

132   inducing long branch attraction (LBA) artefacts; and (2) we wanted to investigate the effects of

133   removing eukaryotes and Korarchaeota, which were the sister lineages of Njordarchaeales in the

134   NM57 and RP56 phylogenetic analyses, respectively. Following this, we recoded the alignments

135   into 4 states (using SR4-recoding[27]) to ameliorate potential phylogenetic artefacts arising from

136   model misspecification at mutationally saturated or compositionally biased sites[14,28–30]. Further,

137   with a similar goal, we applied a fast-evolving site removal (FSR) procedure to the concatenated

138   datasets, since fast-evolving sites are often mutationally saturated. We performed phylogenetic

139   analyses of the above-mentioned datasets in both ML and Bayesian Inference (BI) frameworks,

140   under sophisticated evolutionary models that account for sequence heterogeneity in the

141   substitution process across sites (mixture models; Supplementary Table 2).

142   Phylogenomic analyses of the above-mentioned combinations of taxon sampling, data treatments

143   and phylogenetic frameworks revealed that Njordarchaeales are artefactually attracted to

144   Korarchaeota in RP56 datasets (Supplementary Information). This attraction is likely caused by

145   the high compositional similarity of njord- and korarchaeal RP56 ribosomal protein sequences,

146   which is probably linked to their shared hyperthermophilic lifestyle (Supplementary Figures 4-6).

147   Analyses of RP56 datasets from which Korarchaeota were removed, recovered Njordarchaeales as

7

148   an order at the base of or within Heimdallarchaeia (Supplementary Figure 7), consistent with

149   phylogenomic analyses of the NM57 dataset that included Korarchaeota (Supplementary Figure

150   3). Next, in our efforts to resolve the phylogenetic placement of eukaryotes, we initially performed

151   phylogenomic analyses on variations of the RP56 and NM57 datasets (Supplementary Table 2 and

152   Discussion). However, since compared to the RP56 dataset, the NM57 dataset is larger and less

153   compositionally biased, and is thus expected to have retained a stronger and more congruent

154   phylogenetic signal, we focused the rest of our study on this more reliable dataset.

155

156   **Eukarya represent a well-nested clade within Heimdallarchaeia**

157   Subsequent phylogenetic analyses of untreated NM57 datasets with various taxon sampling

158   variations recovered eukaryotes as sister-clade to Njordarchaeales in ML analyses (e.g.,

159   Supplementary Figure 3, Supplementary Table 2 and Supplementary information). However, ML

160   analyses of the SR4-recoded datasets retrieved a complex phylogenetic signal, as in some cases

161   eukaryotes were placed at the base of all Heimdallarchaeia (including Njordarchaeales) and

162   Wukongarchaeia. This strongly suggests that the previously observed phylogenetic affiliation

163   between Njordarchaeales and eukaryotes could represent an artefact. Furthermore, when both SR4-

164   recoding and FSR treatments were combined, eukaryotes were nested within Heimdallarchaeia, as

165   sister-group to the order Hodarchaeales (Figure 2; Supplementary Figure 8), and this position was

166   supported by ML analyses of NM57 datasets across all taxon selection variations (removing

167   DPANN archaea, and/or Korarchaeota and/or Njordarchaeales). Congruently, the monophyly of

168   eukaryotes and Hodarchaeales was systematically recovered by BI of recoded datasets (both with

169   and without FSR; Figure 2, Supplementary Table 2). In addition, the position of Njordarchaeales

170   shifted during these analyses, moving from a deep position at the base of Heimdallarchaeia and

171 Wukongarchaeia, to a more nested position forming a clade with Gerdarchaeales and Kari- and

172 Heimdallarchaeaceae (Supplementary Discussion). This shift is observed in both the NM57 and

173 the RP56 datasets analyses when SR4-recoding and FSR was combined (Supplementary Figures

174 9-10), supporting that Njordarchaeales represent a divergent order-level lineage of

175 Heimdallarchaeia.

176 In summary, resolving the position of eukaryotes relative to Asgard archaea is anything but trivial

177 (see Supplementary Discussion). In our efforts to extract the true phylogenetic signal, we provide

178 confident support for eukaryotes forming a well-nested clade within the Asgard archaea phylum,

179 consistent with the 2D tree of life scenario. More specifically, we observe that eukaryotes affiliate

180 with the Heimdallarchaeia in analyses in which we systematically reduce phylogenetic artefacts,

181 predominantly converging on a position of eukaryotes as sister to Hodarchaeales, which is also in

182 line with the observed ESP content and genome evolution dynamics (see below).

183

Informational ESPs in Hodarchaeales

185 We found that most of the ESPs previously identified in a limited sampling of Asgard archaea[9,10]

186 are widespread across all phyla included in the present study (Figure 3, Supplementary Table 3).

187 Notably, we observed some exceptions in support of the phylogenetic affiliation between

188 Hodarchaeales and eukaryotes, particularly among ESPs involved in information processing: (1)

189 the ε DNA polymerase subunit is only found in Hodarchaeales; (2) ribosomal protein L28e

190 (Rpl28e/Mak16) homologs are unique to Njord- and Hodarchaeales members; (3) many archaea

191 that lack genes coding for the synthesis of diphthamide, a modified histidine residue which is

192 uniquely present in archaeal and eukaryotic elongation factor 2 (EF-2), instead encode a second

193 EF-2 paralog that misses key-residues required for diphthamide modification[31]. Interestingly, we

194   found that among all Asgard archaea, only MAGs of all sampled Hodarchaeales members encode

195   *dph* genes in addition to a single gene encoding canonical EF-2, which branches at the base of their

196   eukaryotic counterparts in phylogenetic analyses (Supplementary Figure 11; Supplementary

197   Information); (4) While RPL22e and RNA polymerase subunit RPB8 are found in several Asgard

198   archaeal phyla, the only Heimdallarchaeia genomes encoding these genes are members of the

199   Hodarchaeales. Finally, (5) we identified N-terminal histone tails characteristic of eukaryotic

200   histones in all three Hodarchaeales MAGs, as well as in three Njordarchaeales genomes (see

201   Supplementary Information). Altogether, the identification of these key-informational ESPs, in

202   agreement with phylogenomic analyses described above, supports that Hodarchaeales represent

203   the closest archaeal relatives of eukaryotes.

204

Expanded eukaryotic-like protein translocation repertoire

205

206   In our search for putative new ESPs in the expanded Asgard archaeal genomic diversity, we

207   uncovered several additional homologs of proteins associated with the eukaryotic translocon, a

208   protein complex primarily responsible for the post-translational modification of proteins, and

209   subsequent insertion into, or transport across the membrane of the endoplasmic reticulum (ER)[32].

210   The eukaryotic translocon is comprised of the core Sec61 protein-conducting channel, and several

211   accessory components, including the oligosaccharyltransferase (OST) and translocon-associated

212   protein (TRAP) complexes (Figure 3b), both of which are involved in the biogenesis of N-

213   glycosylated proteins[33]. The eukaryotic TRAP complex is composed of two to four subunits in

214   eukaryotes. Using distant-homology detection methods, we identified homologs from three of

215   these subunits to be broadly distributed across Asgard archaeal genomes, while the fourth one was

216   detected only in a few thorarchaeial MAGs (Figure 3b). The eukaryotic OST complex generally

10

217 comprises 6-8 subunits organized into three subcomplexes that are collectively embedded in the

218 ER membrane[34] (Figure 3b). Apart from STT3/AglB (OST subcomplex-II), which represents the

219 catalytic subunit and is universally found across all three domains of life, other OST subcomplexes

220 generally do not possess prokaryotic homologs beyond the Ost1/Ribophorin I (OST subcomplex-

221 I) and Ost3/Tusc3 (OST subcomplex-II) subunits previously reported in Asgard archaea[10]. Here,

222 we report the identification of Asgard archaeal homologs of all five additional subunits,

223 Ost2/Dad1, Ost4, Ost5/TMEM258, SWP1/Ribophorin II and WBP1/Ost48. While we identified

224 homologs of Ost4 and Ost5 (OST subcomplex-I) in most Asgard archaeal classes, the distribution

225 of Ost2, WBP1, and Swp1, the first subcomplex-III subunits described in prokaryotes to date, was

226 restricted to Heimdallarchaeia, including Njordarchaeales for WBP1, further supporting their

227 monophyly. Our findings indicate that Asgard archaea and, by inference, LAECA, potentially

228 encode relatively complex machineries for the N-linked glycosylation and translocation of proteins

229 (Figure 3b).

230

231 **Vesicular biogenesis and trafficking proteins**

232 Intracellular vesicular transport represents a key process that emerged during eukaryogenesis.

233 Previous studies have reported that Asgard archaeal genomes encode homologs of eukaryotic

234 proteins comprising various intracellular vesicular trafficking and secretion machineries, including

235 the ESCRT (endosomal sorting complexes required for transport), TRAPP (transport protein

236 particle) and COPII (coat protein complex II) vesicle coatomer protein complexes[9,10]. Furthermore,

237 as much as 2% of the genes of Asgard archaeal genomes were found to encode small GTPase

238 homologs – a broad family of eukaryotic proteins, encompassing the Ras, Rab, Arf, Rho and Ran

239 subfamilies, that are broadly implicated in budding, transport, docking and fusion of vesicles in

240   eukaryotic cells[9,10]. Here, we report the identification of Asgard archaeal homologs of subunits of

241   additional vesicular trafficking complexes (Figure 3, Ext. Data Fig. 2, Supplementary Table 3).

242   Noticeably, we found putative homologs of all four subunits composing eukaryotic adaptor

243   proteins (AP) and coatomer protein (COPI) complexes, which, in eukaryotic cells, are involved in

244   the formation of clathrin-coated pits and vesicles responsible for packaging and sorting cargo for

245   transport through the secretory and endocytic pathways[35]. Those complexes are composed of two

246   large subunits, belonging to the β- and γ-families, a medium μ-subunit, and a small σ-subunit. We

247   found homologs of all functional domains composing those subunits, albeit sparsely distributed

248   (Ext. Data Fig. 2, Supplementary Information). Additionally, we found homologs of several

249   protein complexes involved in eukaryotic endosomal sorting such as the retromer, the

250   HOPS/CORVET and the GARP complexes (Figure 3, red shading). Retromer is a coat-like

251   complex associated with endosome-to-Golgi retrograde traffic[36] and we detected four of its five

252   subunits in Asgard MAGs. One of these subunits is Vps5-BAR, which in Thorarchaeia is often

253   fused to Vps28, a subunit of the ESCRT-I subcomplex, suggesting a functional link between BAR

254   domain proteins and the thorarchaeial ESCRT complex. The GARP (Golgi-associated retrograde

255   protein) complex is a multisubunit tethering complex located at the trans-Golgi network in

256   eukaryotic cells, where it also functions to tether retrograde transport vesicles derived from

257   endosomes[37,38], similarly to the retromer. GARP comprises four subunits, three of which we could

258   detect in Asgard archaeal genomes, with a sparse and punctuated distribution. Functioning in

259   opposite direction from the retromer and GARP complexes are the CORVET (Class C core

260   vacuole/endosome tethering) and HOPS (Homotypic fusion and protein sorting) complexes[39].

261   Endosomal fusion and autophagy in eukaryotic cells depend on them and they share four core

12

262   subunits[40], three of which can be found in Asgard archaea, in addition to one of the HOPS specific

263   subunits[41].

264   Finally, while numerous components of the ESCRT-I, II and III systems have been previously

265   detected in Asgard archaea[9,10,42], we report here the identification of Asgard homologs for the

266   ESCRT-III regulators Vfa1, Vta1, Ist1, and Bro1.

267

268   **Ancestral Asgard archaea genome reconstruction**

269   The analysis of Asgard archaeal genome data obtained through metagenomics, combined with the

270   insights derived from cytological observations of the first two cultured Asgard archaea

271   'Candidatus Prometheoarchaeum syntrophicum'[43] and 'Candidatus Lokiarchaeum ossiferum'[44],

272   have generated new hypotheses about the nature of the archaeal ancestor of eukaryotes[43,45–48].

273   However, these theories are mostly based on a limited number of features displayed by a single,

274   or a few Asgard archaeal lineages. While informative, features of present-day Asgard archaea do

275   not necessarily resemble those of LAECA, as these are potentially separated by over 2 Gya of

276   evolution[49]. Furthermore, Asgard archaeal phyla display a highly variable genome content with

277   respect to ESPs and predicted metabolic features[43,46,48,50,51], suggesting a complex evolutionary

278   history of those traits. In light of these considerations, we inferred ancestral features of LAECA

279   by using an ML evolutionary framework. We employed a recently developed probabilistic gene-

280   tree species-tree reconciliation approach[52,53] in combination with the extended taxonomic

281   sampling of Asgard archaeal genomes to reconstruct the evolutionary history of homologous gene

282   families and ancestral gene content across the Asgard archaeal species tree. For this, we inferred

283   ML phylogenetic trees of all 17,200 protein families encoded across 181 archaeal genomes,

284   including representatives from Asgard and TACK archaea, and Euryarchaeota clades. Importantly,

285   as missing genes and potential contaminations in MAGs will be regarded as recent gene loss and

286   gain events in our ancestral reconstruction analyses, the use of incomplete MAGs with low

287   contamination levels is unlikely to have a major impact on the inferred gene content of the deep

288   archaeal ancestors that were reconstructed in the present study (also see Supplementary

289   Information).

290   We first compared the distributions of estimated ancestral proteome sizes, and numbers of inferred

291   gene duplications, losses and gains (i.e., horizontal gene transfers and originations) in all archaeal

292   ancestral nodes (Supplementary Figure 12). Intriguingly, we observed that Heimdall- (and

293   particularly the ancestor of Hodarchaeales) and Lokiarchaeia ancestors display significantly higher

294   gene duplication rates compared to TACK and Euryarchaeota ancestors (Figure 4a). In addition,

295   we found that most Asgard archaeal ancestors displayed gene loss rates comparable to other

296   archaea, with the exception of Thorarchaeia, Lokiarchaeales and Jordarchaeia, which showed

297   significantly lower loss rates. In agreement with the observed evolutionary genome dynamics, we

298   found that predicted proteome sizes of most Asgard archaea ancestors are significantly larger than

299   other archaeal ancestors (P<0.001), with Lokiarchaeia ancestors displaying the largest estimated

300   proteome size (Supplementary Figure 13). Similarly, the Hodarchaeales ancestor had an estimated

301   proteome size of 4,053 proteins, versus 3,134 for the last Asgard archaea common ancestor

302   (LAsCA), reflecting the high duplication and low loss rates in that clade. The streamlined genome

303   content of the Odinarchaeia ancestor represents an exception to the general trend of genome

304   expansion across Asgard archaea, and possibly reflects an adaptation to high temperatures[54].

305

Ancestral features of LAECA

Using the approach described above, we also reconstructed the ancestral metabolic and physiological properties across the Asgard archaeal species tree, including the proposed closest archaeal relatives of eukaryotes, the Hodarchaeales. We infer that the LAsCA was a chemolithotroph that required the synthesis of organic building blocks via the Wood-Ljungdahl pathway (WLP) (Figure 4b and Supplementary information), for which we inferred the presence of key enzymes, including carbon monoxide dehydrogenase/acetyl-CoA synthase (CODH/ACS) and the formylmethanofuran dehydrogenase (FmdABCDE). In addition, our analyses revealed that the last common ancestors of individual Asgard archaeal phyla either had the genetic potential to switch between autotrophy and heterotrophy (Loki-, Thor-, Jord- and Baldrarchaeia) or a predominantly heterotrophic fermentative (Odin- and Heimdallarchaeia) lifestyle (Figure 4b, Supplementary Information). Specifically, we observed that the WLP was lost prior to the split between Njordarchaeales and the other Heimdallarchaeia (and therefore prior to the emergence of LAECA), indicating that LAECA was a heterotrophic fermenter (Supplementary Table 4).

Furthermore, we infer that the central carbon metabolism of Heimdallarchaeia (including Hodarchaeales) included the Embden-Meyerhof-Parnas (EMP) pathway and a partial oxidative pentose phosphate (OPP) pathway - both considered core modules of present-day eukaryotic central carbon metabolism. While the enzymes of these pathways in Asgard archaea do not share a common evolutionary origin with those of eukaryotes, this indicates that LAECA had a similar central carbon metabolism compared to modern eukaryotes (Supplementary Figure 14-15).

In addition, our analyses support the idea that the last common ancestor of Heimdallarchaeia contained several components of the electron transport chain (ETC)[48]. We inferred that the last common ancestor of Hodarchaeales likely contained CI, CII, CIV and a nitrate reductase complex

15

329   (NarGHIJ), indicating that nitrate might have been used as a terminal electron acceptor to perform

330   anaerobic respiration. As such, the last Hodarchaeales common ancestor likely generated ATP

331   using an electron transport chain where electrons from NADH and succinate were transferred

332   through a series of membrane-associated complexes with quinones and cupredoxins as electron

333   carriers to ultimately reduce nitrate[55].

334   As indicated above, a significant fraction of the currently sampled Asgard archaea diversity

335   originates from geothermal or hydrothermal environments. Indeed, using an algorithm based on

336   genome-derived features[56], we confirmed that (most) Njordarchaeales, Baldr- and Jordarchaeia are

337   hyperthermophiles, Odinarchaeia are thermophiles, and Loki- and Thorarchaeia are mesophiles

338   (Figure 4c, Supplementary Table 5). While Heimdallarchaeia seem to contain both meso- and

339   thermophiles, we infer a mesophilic physiology for Hodarchaeales, obtaining the lowest predicted

340   optimal growth temperatures among all Asgard archaea (median=36.7 °C). Asgard archaeal

341   hyperthermophiles contain reverse gyrase, a topoisomerase that is typically encoded by

342   hyperthermophilic prokaryotes[57]. We infer that a reverse gyrase was possibly present in LAsCA

343   and that it was subsequently lost in all heimdallarchaeial orders except for Njordarchaeales. This

344   observation would be compatible with a scenario in which Asgard archaea have a

345   hyperthermophilic ancestry, but in which eukaryotes evolved from an Asgard archaea lineage that

346   had adapted to mesophilic growth temperatures.

347

348   **Discussion**

349   Beyond genomic exploration, several studies have started to unveil important physiological,

350   cytological and ecological aspects of Asgard archaea[43,58–60]. Yet, while such insights are certainly

351   relevant, the cellular and physiological characteristics of present-day Asgard archaea will almost

16

352   certainly not resemble those of LAECA. Therefore, inferences about the identity and nature of

353   LAECA and the process of eukaryogenesis should be made within an evolutionary context. We

354   used an evolutionary framework to analyze an expanded Asgard archaeal genomic diversity,

355   comprising 11 clades of high taxonomic rank. Using comprehensive phylogenomic analyses

356   involving the evaluation of distinct marker protein datasets and systematic assessment of suspected

357   phylogenetic artefacts and state-of-the-art models of evolution, we identified Hodarchaeales, a

358   class-level clade within the Heimdallarchaeia, as the closest relatives of eukaryotes. Evidently,

359   phylogenomic analyses aiming to pinpoint the phylogenetic position of eukaryotes in the tree of

360   life are extremely challenging, and our results stress the importance of testing for possible sources

361   of bias affecting phylogenomic reconstructions, as was recently reviewed[61]. The implementation

362   of a probabilistic gene tree/species tree reconciliation approach allowed us to infer the evolutionary

363   dynamics and ancestral content across the archaeal species tree providing several new insights into

364   the Asgard archaeal roots of eukaryotes. Altogether, our results reveal a picture in which the

365   Asgard archaeal ancestor of eukaryotes had, compared to other archaea, a relatively large genome,

366   resulting mainly from more numerous gene duplication and fewer gene loss events. It is tempting

367   to speculate that the elevated gene duplication rates observed in our analyses represent an ancestral

368   feature of LAECA, and that it remained the predominant modus of genome evolution during the

369   early stages of eukaryogenesis. We also inferred that the duplicated gene content of LAECA

370   included several protein families involved in cytoskeletal and membrane-trafficking functions,

371   including among others actin homologs, ESCRT complex subunits and small GTPase homologs.

372   Our findings complement those of another study[62] reporting that eukaryotic proteins with an

373   Asgard archaeal provenance, as opposed to those inherited from the mitochondrial symbiont,

17

374   duplicated the most during eukaryogenesis, particularly proteins of cytoskeletal and membrane-

375   trafficking families.

376   Beyond genome dynamics, our analyses of inferred ancestral genome content across the Asgard

377   archaeal species tree indicates that, while Asgard archaea likely had a thermophilic ancestry, the

378   lineage from which eukaryotes evolved was adapted to mesophilic conditions, which is compatible

379   with a generally assumed mesophilic ancestry of eukaryotes. Furthermore, we infer that LAECA

380   had the genetic potential to support a heterotrophic lifestyle, and may have been able to conserve

381   energy via nitrate respiration. In addition, based on taxonomic distribution and evolutionary

382   history of ESPs we show that complex pathways involved in protein targeting and membrane

383   trafficking, and in genome maintenance and expression in eukaryotes were inherited from their

384   Asgard archaeal ancestor. Of note, we identified additional Asgard archaeal homologs of

385   eukaryotic vesicular trafficking complex components. Of these, some Asgard archaeal proteins

386   display sequence similarity to proteins which, in eukaryotes, are part of the clathrin adaptor protein

387   complexes and of the COPI complex. These complexes are particularly interesting since they are

388   involved in the biogenesis of vesicles responsible for sorting cargo and subsequent transport

389   through the secretory and endocytic pathways[35]. Altogether, these results further suggest a

390   potential for membrane deformation, and possibly trafficking, in Asgard archaea. The ability to

391   deform membranes was recently shown in two papers reporting the first cultivated Lokiarchaeia

392   lineages, '*Ca.* P. syntrophicum strain MK-D1'[43] and '*Ca.* Lokiarchaeum ossiferum'[44], whose cells

393   both displayed unique morphological complexity including long and often branching protrusions

394   facilitated by a dynamic actin cytoskeleton. Thus far no[43], or only limited[44] visible endomembrane

395   structures were observed in these first cultured representatives of Asgard archaea. However, it is

396   important to restate here that, being separated by some 2 Gya of evolution, the cellular features of

18

397   present-day Asgard archaeal lineages do not necessarily resemble those of LAECA. Furthermore,

398   given the disparity of the distribution patterns of membrane trafficking homologs in Asgard

399   archaea, it will be crucial to isolate representatives of phyla other than Lokiarchaeia and study their

400   cell biological features and potential for endomembrane biogenesis. Of particular interest would

401   be members of the Heimdallarchaeia, and specifically Hodarchaeales, as the currently identified

402   closest relatives of eukaryotes, as well as Thorarchaeia lineages, which seem to generally contain

403   a particularly rich suite of homologs of eukaryotic membrane trafficking proteins.

404   By phylogenetically placing eukaryotes as a firmly-nested clade within the presently identified

405   Asgard archaeal diversity, and by inferring ancestral genomic content across the Asgard archaea,

406   our work provides insights into the identity and nature of the Asgard archaeal ancestor of

407   eukaryotes, guiding future studies aiming to uncover new pieces of the elusive eukaryogenesis

408   puzzle.

409

References

411   1.   López-García, P. & Moreira, D. Open Questions on the Origin of Eukaryotes. *Trends Ecol. Evol.* **30**,

412   697–708 (2015).

413   2.   Eme, L., Spang, A., Lombard, J., Stairs, C. W. C. W. & Ettema, T. J. G. T. J. G. Archaea and the

414   origin of eukaryotes. *Nat. Rev. Microbiol.* **15**, nrmicro–2017 (2017).

415   3.   Guy, L. & Ettema, T. J. G. The archaeal 'TACK' superphylum and the origin of eukaryotes. *Trends*

416   *Microbiol.* **19**, 580–587 (2011).

417   4.   Kelly, S., Wickstead, B. & Gull, K. Archaeal phylogenomics provides evidence in support of a

418   methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc. Biol. Sci.*

419   **278**, 1009–1018 (2011).

420 5. Williams, T. a., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports

421   only two primary domains of life. *Nature* **504**, 231–236 (2013).

422 6. Lasek-Nesselquist, E. & Gogarten, J. P. The effects of model choice and mitigating bias on the

423   ribosomal tree of life. *Mol. Phylogenet. Evol.* (2013) doi:10.1016/j.ympev.2013.05.006.

424 7. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new

425   root for the Archaea. *Proceedings of the National Academy of Sciences* 201420858 (2015).

426 8. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaebacterial origin of

427   eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20356–20361 (2008).

428 9. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*

429   **521**, 173–179 (2015).

430 10. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular

431   complexity. *Nature* **541**, 353–358 (2017).

432 11. Cunha, V. D. *et al.* Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between

433   prokaryotes and eukaryotes. *PLOS Genetics* vol. 13 e1006810 Preprint at

434   https://doi.org/10.1371/journal.pgen.1006810 (2017).

435 12. Da Cunha, V., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the debate about the

436   universal tree of life topology. *PLoS genetics* vol. 14 e1007215 (2018).

437 13. Spang, A. *et al.* Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet.* **14**,

438   (2018).

439 14. Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J. & Embley, T. M. Phylogenomics provides

440   robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

441 15. Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature*

442   **593**, 553–557 (2021).

443 16. Hug, L. A. *et al.* Community genomic analyses constrain the distribution of metabolic traits across

444   the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**, 22 (2013).

445 17. Zhang, J.-W. *et al.* Newly discovered Asgard archaea Hermodarchaeota potentially degrade alkanes

446   and aromatics via alkyl/benzyl-succinate synthase and benzoyl-CoA pathway. *ISME J.* (2021)

447   doi:10.1038/s41396-020-00890-x.

448   18. Farag, I. F., Zhao, R. & Biddle, J. F. 'Sifarchaeota' a novel Asgard phylum from Costa Rica

449   sediment capable of polysaccharide degradation and anaerobic methylotrophy. *Appl. Environ.*

450   *Microbiol.* (2021).

451   19. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-

452   assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).

453   20. Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in

454   the surface ocean. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11463–11468 (2013).

455   21. Sun, J. *et al.* Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota

456   lineages. *ISME Communications* **1**, 1–14 (2021).

457   22. Xie, R. *et al.* Expanding Asgard members in the domain of Archaea sheds new light on the origin of

458   eukaryotes. *Sci. China Life Sci.* **65**, 818–829 (2022).

459   23. Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? *Mol.*

460   *Phylogenet. Evol.* **75**, 103–117 (2014).

461   24. Petitjean, C., Deschamps, P., López-García, P. & Moreira, D. Rooting the Domain Archaea by

462   Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. *Genome*

463   *Biol. Evol.* **7**, 191–204 (2014).

464   25. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. Origins of coevolution between

465   residues distant in protein 3D structures. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9122–9127 (2017).

466   26. Foster, P. G. & Hickey, D. A. Compositional bias may affect both DNA-based and protein-based

467   phylogenetic reconstructions. *J. Mol. Evol.* **48**, (1999).

468   27. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol.*

469   *Evol.* **24**, 2139–2150 (2007).

470   28. Brown, M. W. *et al.* Phylogenomics demonstrates that breviate flagellates are related to opisthokonts

471   and apusomonads. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20131755–

21

472   20131755 (2013).

473   29. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of

474   incongruence? *Trends in Genetics* vol. 22 225–231 Preprint at

475   https://doi.org/10.1016/j.tig.2006.02.003 (2006).

476   30. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent Genome Reduction and

477   Phylogenetic Reclassification of the Oceanic SAR11 Clade. *Molecular Biology and Evolution* vol.

478   29 599–615 Preprint at https://doi.org/10.1093/molbev/msr203 (2012).

479   31. Narrowe, A. B. *et al.* Complex Evolutionary History of Translation Elongation Factor 2 and

480   Diphthamide Biosynthesis in Archaea and Parabasalids. *Genome Biol. Evol.* **10**, 2380–2393 (2018).

481   32. Wang, L. & Dobberstein, B. Oligomeric complexes involved in translocation of proteins across the

482   membrane of the endoplasmic reticulum. *FEBS Lett.* **457**, 316–322 (1999).

483   33. Pfeffer, S. *et al.* Dissecting the molecular organization of the translocon-associated protein complex.

484   *Nat. Commun.* **8**, 14516 (2017).

485   34. Bai, L., Wang, T., Zhao, G., Kovach, A. & Li, H. The atomic structure of a eukaryotic

486   oligosaccharyltransferase complex. *Nature* **555**, 328–333 (2018).

487   35. Rout, M. P. & Field, M. C. The Evolution of Organellar Coat Complexes and Organization of the

488   Eukaryotic Cell. *Annu. Rev. Biochem.* **86**, 637–657 (2017).

489   36. Seaman, M. N. J. The retromer complex - endosomal protein recycling and beyond. *J. Cell Sci.* **125**,

490   4693–4702 (2012).

491   37. Liewen, H. *et al.* Characterization of the human GARP (Golgi associated retrograde protein)

492   complex. *Exp. Cell Res.* **306**, 24–34 (2005).

493   38. Pérez-Victoria, F. J. *et al.* Structural basis for the wobbler mouse neurodegenerative disorder caused

494   by mutation in the Vps54 subunit of the GARP complex. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12860–

495   12865 (2010).

496   39. Villaseñor, R., Kalaidzidis, Y. & Zerial, M. Signal processing by the endosomal system. *Curr. Opin.*

497   *Cell Biol.* **39**, 53–60 (2016).

22

498   40. Graham, S. C. *et al.* Structural basis of Vps33A recruitment to the human HOPS complex by Vps16.

499   *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13345–13350 (2013).

500   41. Jiang, P. *et al.* The HOPS complex mediates autophagosome–lysosome fusion through interaction

501   with syntaxin 17. *Molecular Biology of the Cell* vol. 25 1327–1337 Preprint at

502   https://doi.org/10.1091/mbc.e13-08-0447 (2014).

503   42. Hatano, T. *et al.* Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-

504   ESCRT machinery. *Nat. Commun.* **13**, 3398 (2022).

505   43. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525

506   (2020).

507   44. Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard

508   archaeon. *Nature* Preprint at https://doi.org/10.1038/s41586-022-05550-y (2022).

509   45. Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N. & Martin, W. F. Lokiarchaeon is hydrogen

510   dependent. *Nat Microbiol* **1**, 16034 (2016).

511   46. Bulzu, P.-A. *et al.* Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nature*

512   *Microbiology* vol. 4 1129–1137 Preprint at https://doi.org/10.1038/s41564-019-0404-y (2019).

513   47. López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat*

514   *Microbiol* **5**, 655–667 (2020).

515   48. Spang, A. *et al.* Proposal of the reverse flow model for the origin of the eukaryotic cell based on

516   comparative analyses of Asgard archaeal metabolism. *Nat Microbiol* **4**, 1138–1148 (2019).

517   49. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and

518   eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).

519   50. Seitz, K. W. *et al.* Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat. Commun.* **10**,

520   1822 (2019).

521   51. Liu, Y. *et al.* Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota.

522   *ISME J.* **12**, 1021–1031 (2018).

523   52. Szöllõsi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the

23

524   space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).

525 53. Szöllősi, G. J., Davín, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-scale phylogenetic

526   analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**,

527   20140335 (2015).

528 54. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth Temperature and Genome Size in Bacteria

529   Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. *Genome*

530   *Biol. Evol.* **5**, 966–977 (2013).

531 55. Savelieff, M. G. *et al.* Experimental evidence for a link among cupredoxins: red, blue, and purple

532   copper transformations in nitrous oxide reductase. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7919–7924

533   (2008).

534 56. Sauer, D. B. & Wang, D.-N. Predicting the optimal growth temperatures of prokaryotes using only

535   genome derived features. *Bioinformatics* Preprint at https://doi.org/10.1093/bioinformatics/btz059

536   (2019).

537 57. Pavel Lulchev, D. K. Reverse gyrase—recent advances and current mechanistic understanding of

538   positive DNA supercoiling. *Nucleic Acids Res.* **42**, 8200 (2014).

539 58. Akıl, C. & Robinson, R. C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature*

540   **562**, 439–443 (2018).

541 59. Orsi, W. D. *et al.* Metabolic activity analyses demonstrate that Lokiarchaeon exhibits

542   homoacetogenesis in sulfidic marine sediments. *Nat Microbiol* **5**, 248–255 (2020).

543 60. Neveu, E., Khalifeh, D., Salamin, N. & Fasshauer, D. Prototypic SNARE Proteins Are Encoded in

544   the Genomes of Heimdallarchaeota, Potentially Bridging the Gap between the Prokaryotes and

545   Eukaryotes. *Curr. Biol.* **30**, 2468–2480.e5 (2020).

546 61. Williams, T. A. *et al.* Inferring the deep past from molecular data. *Genome Biol. Evol.* (2021)

547   doi:10.1093/gbe/evab067.

548 62. Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications.

549   *Nat Ecol Evol* **5**, 92–100 (2021).

550   63.   Hua, Z.-S. *et al.* Genomic inference of the metabolism and evolution of the archaeal phylum

551   Aigarchaeota. *Nat. Commun.* **9**, 2832 (2018).

552   64.   Chen, L.-X. *et al.* Candidate Phyla Radiation Roizmanbacteria From Hot Springs Have Novel and

553   Unexpectedly Abundant CRISPR-Cas Systems. *Front. Microbiol.* **10**, 928 (2019).

554   65.   Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–

555   359 (2012).

556   66.   Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately

557   reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

558   67.   Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.*

559   **10**, R85 (2009).

560   68.   Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria.

561   *Nature* **523**, 208–211 (2015).

562   69.   Flores, G. E. *et al.* Inter-field variability in the microbial communities of hydrothermal vent deposits

563   from a back-arc basin. *Geobiology* **10**, 333–346 (2012).

564   70.   Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data.

565   *Bioinformatics* **30**, 2114–2120 (2014).

566   71.   Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis.

567   *F1000Res.* **4**, 900 (2015).

568   72.   Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node

569   solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*

570   vol. 31 1674–1676 Preprint at https://doi.org/10.1093/bioinformatics/btv033 (2015).

571   73.   Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced

572   methodologies and community practices. *Methods* **102**, 3–11 (2016).

573   74.   Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of

574   short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

575   75.   Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079

25

576   (2009).

577   76.   Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the

578   quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*

579   **25**, 1043–1055 (2015).

580   77.   Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a

581   novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur

582   reduction. *ISME J.* **10**, 1696–1705 (2016).

583   78.   Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**,

584   1144–1146 (2014).

585   79.   Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of

586   metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**, 236 (2015).

587   80.   Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell

588   sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).

589   81.   Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-

590   cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* vol. 28 1420–1428

591   Preprint at https://doi.org/10.1093/bioinformatics/bts174 (2012).

592   82.   Hugoson, E., Lam, W. T. & Guy, L. miComplete: weighted quality evaluation of assembled

593   microbial genomes. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz664.

594   83.   Karst, S. M., Kirkegaard, R. H. & Albertsen, M. mmgenome: a toolbox for reproducible genome

595   extraction from metagenomes. Preprint at https://doi.org/10.1101/059121.

596   84.   Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage

597   binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).

598   85.   Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and

599   biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9**, 4999 (2018).

600   86.   Huang, J.-M., Baker, B. J., Li, J.-T. & Wang, Y. New Microbial Lineages Capable of Carbon

601   Fixation and Nutrient Cycling in Deep-Sea Sediments of the Northern South China Sea. *Appl.*

602   *Environ. Microbiol.* **85**, (2019).

603   87. Tréhu, A. M. *et al.* Feeding methane vents and gas hydrate deposits at south Hydrate Ridge.

604   *Geophysical Research Letters* vol. 31 Preprint at https://doi.org/10.1029/2004gl021286 (2004).

605   88. Nunoura, T., Inagaki, F., Delwiche, M. E., Colwell, F. S. & Takai, K. Subseafloor microbial

606   communities in methane hydrate-bearing sediment at two distinct locations (ODP Leg204) in the

607   cascadia margin. *Microbes Environ.* **23**, 317–325 (2008).

608   89. Inagaki, F. *et al.* Biogeographical distribution and diversity of microbes in methane hydrate-bearing

609   deep marine sediments on the Pacific Ocean Margin. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2815–2820

610   (2006).

611   90. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads.

612   *EMBnet.journal* vol. 17 10 Preprint at https://doi.org/10.14806/ej.17.1.200 (2011).

613   91. Dombrowski, N., Seitz, K. W., Teske, A. P. & Baker, B. J. Genomic insights into potential

614   interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments.

615   *Microbiome* **5**, 106 (2017).

616   92. Parks, D. H. *et al.* Author Correction: A complete domain-to-species taxonomy for Bacteria and

617   Archaea. *Nat. Biotechnol.* **38**, 1098 (2020).

618   93. Romalde, J. L., Balboa, S. & Ventosa, A. *Microbial Taxonomy, Phylogeny and Biodiversity.*

619   (Frontiers Media SA, 2019).

620   94. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior

621   Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* **67**, 216–

622   235 (2018).

623   95. Lagkouvardos, I. *et al.* IMNGS: A comprehensive open resource of processed 16S rRNA microbial

624   profiles for ecology and diversity studies. *Sci. Rep.* **6**, 33721 (2016).

625   96. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* vol. 4 1686

626   Preprint at https://doi.org/10.21105/joss.01686 (2019).

627   97. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

27

628   98. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. in

629   *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 1–14 (Springer New York, 2019).

630   99. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat.*

631   *Methods* **12**, 59–60 (2015).

632   100. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX.

633   *BMC Bioinformatics* **12**, 116 (2011).

634   101. Miele, V. *et al.* High-quality sequence clustering guided by network topology and multiple

635   alignment likelihood. *Bioinformatics* **28**, 1078–1085 (2012).

636   102. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program.

637   *Brief. Bioinform.* **9**, 286–298 (2008).

638   103. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence

639   searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).

640   104. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional

641   annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–93 (2016).

642   105. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs):

643   An Update and Application for Analysis of Shared Features between Thermococcales,

644   Methanococcales, and Methanobacteriales. *Life* **5**, 818–840 (2015).

645   106. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* gkt1223 (2013).

646   107. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**,

647   1236–1240 (2014).

648   108. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for

649   Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731

650   (2016).

651   109. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase

652   classification and analysis. *Sci. Rep.* **6**, 34212 (2016).

653   110. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

654 111. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal
655   for protein modeling, prediction and analysis. *Nature Protocols* vol. 10 845–858 Preprint at
656   https://doi.org/10.1038/nprot.2015.053 (2015).

657 112. Petitjean, C., Deschamps, P., López-García, P., Moreira, D. & Brochier-Armanet, C. Extending the
658   conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. *Mol.*
659   *Biol. Evol.* **32**, 1242–1254 (2015).

660 113. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
661   improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

662 114. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software
663   for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol.*
664   *Biol.* **10**, 210 (2010).

665 115. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for
666   large alignments. *PLoS One* **5**, (2010).

667 116. Camacho, C. *et al.* BLAST : architecture and applications. *BMC Bioinformatics* vol. 10 421 Preprint
668   at https://doi.org/10.1186/1471-2105-10-421 (2009).

669 117. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by
670   eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).

671 118. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence
672   similarity searches. *Bioinformatics* **31**, 926–932 (2015).

673 119. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment
674   trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

675 120. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. Iq-tree: A fast and effective
676   stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274
677   (2015).

678 121. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome
679   visualization in R. *Bioinformatics* vol. 26 2334–2335 Preprint at

680   https://doi.org/10.1093/bioinformatics/btq413 (2010).

681   122. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the

682   Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).

683   123. Martijn, J. *et al.* Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile

684   transition. *Nature Communications* vol. 11 Preprint at https://doi.org/10.1038/s41467-020-19200-2

685   (2020).

686   124. Huang, W.-C. *et al.* Comparative genomic analysis reveals metabolic flexibility of Woesearchaeota.

687   *Nat. Commun.* **12**, 5281 (2021).

688   125. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Phylogenetic affiliation of

689   mitochondria with Alpha-II and Rickettsiales is an artefact. *Nature Ecology & Evolution* vol. 6

690   1829–1831 Preprint at https://doi.org/10.1038/s41559-022-01871-3 (2022).

691   126. Dharamshi, J. E. *et al.* Gene gain facilitated endosymbiotic evolution of Chlamydiae. *Nat Microbiol*

692   **8**, 40–54 (2023).

693   127. Kim, E. *et al.* Implication of mouse Vps26b–Vps29–Vps35 retromer complex in sortilin trafficking.

694   *Biochem. Biophys. Res. Commun.* **403**, 167–171 (2010).

695   128. Suzuki, S. W., Chuang, Y.-S., Li, M., Seaman, M. N. J. & Emr, S. D. A bipartite sorting signal

696   ensures specificity of retromer complex in membrane protein recycling. *J. Cell Biol.* **218**, 2876–2886

697   (2019).

698   129. Balderhaar, H. J. K. & Ungermann, C. CORVET and HOPS tethering complexes - coordinators of

699   endosome and lysosome fusion. *J. Cell Sci.* **126**, 1307–1316 (2013).

700
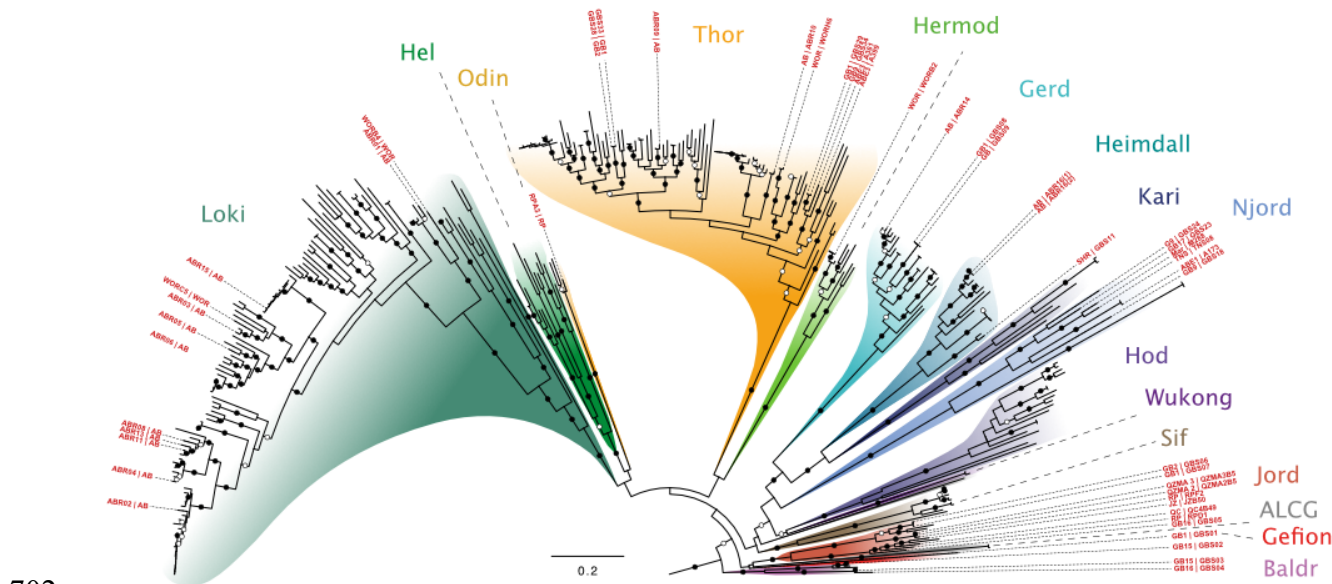
Figures

Figure 1. Phylogenomic analysis of 15 concatenated ribosomal proteins expands Asgard archaea diversity. Maximum-likelihood tree (IQ-TREE, WAG+C60+R4+F+PMSF model) of concatenated protein sequences from at least five genes, encoded on a single contig, of a 15 conserved ribosomal protein (RP15) gene cluster retrieved from publicly available and newly reported Asgard archaeal MAGs. Bootstrap support (100 pseudo-replicates) is indicated by circles at branches, with filled and open circles representing 90% and 70% support, respectively. Leaf names indicate geographical source and isolate name (inner and outer label, respectively) for the MAGs reported in this study. Scale bar denotes the average number of substitutions per site. Abbreviations: AB: Aarhus Bay (Denmark); ABE: ABE vent field, Eastern Lau Spreading Center; ALCG: Asgard Lake Cootharaba Group; CR: Colorado River (USA); GB: Guaymas Basin (Mexico); JZ: Jinze (China); LC: Loki's castle; LCB: Lower Culex Basin (USA); Mar: Mariner vent field, Eastern Lau Spreading Center; NSCS: Northern South China Sea; OWC: Old Woman Creek (USA); QC: QuCai village (China); QZM: QuZhuoMu village (China); RP: Radiata Pool

716   (New Zealand); RS: Red sea; SHR: South Hydrate Ridge; TNS: Taketomi Island (Japan); WOR:

717   White Oak River (USA).

718

719
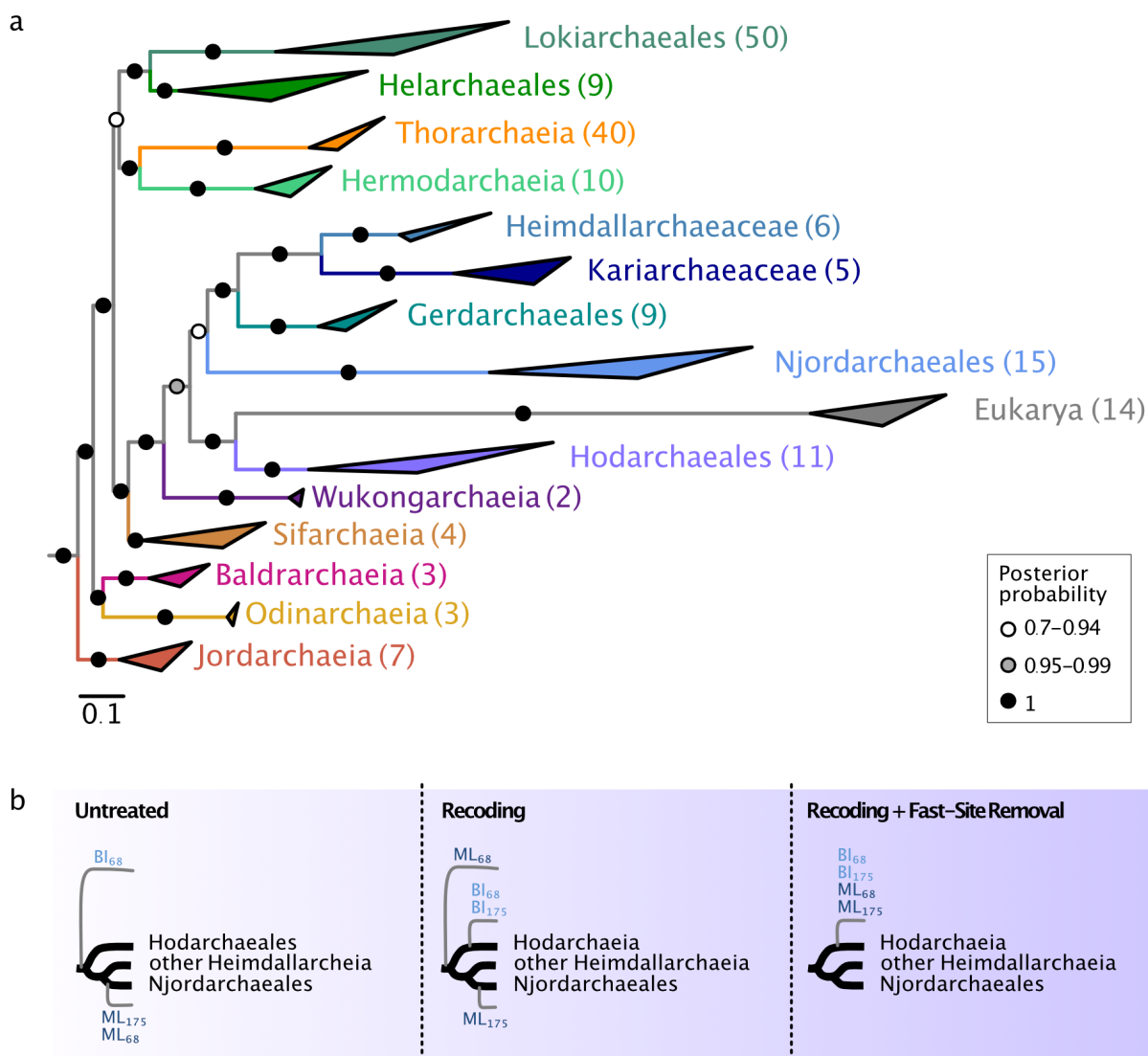
Figure 2. Phylogenomic analyses based on 57 concatenated non-ribosomal proteins support the emergence of eukaryotes as sister to Hodarchaeales. a. Bayesian inference (BI) based on 278 archaeal taxa, using Euryarchaeota and TACK archaea as outgroup (not shown) (NM57-nDK_sr4 alignment, 15,733 amino acid positions). The concatenation was SR4-recoded and

724 analyzed using the CAT+GTR model (4 chains, ~25,000 generations). b. Schematic representation

725 of the shift in the position of eukaryotes (grey branches) in ML and BI analyses of this dataset

726 under different treatments. Untreated: unprocessed dataset; Recoding: SR4-recoded dataset;

727 Recoding+Fast-Site Removal: Fast-site removal combined with SR4-recoding (the topology most

728 often recovered after removing 10% to 50% fastest-evolving sites, in steps of 10%, is shown). 175

729 and 68 refer to phylogenomic datasets containing 175 and 68 Asgard archaea, respectively. Note

730 that BI was not performed for the 175 untreated dataset due to computational limitations (for

731 detailed overview of phylogenomic analyses, see Supplementary Table 3). Scale bar denotes the

732 average expected number of substitutions per site.

733

734
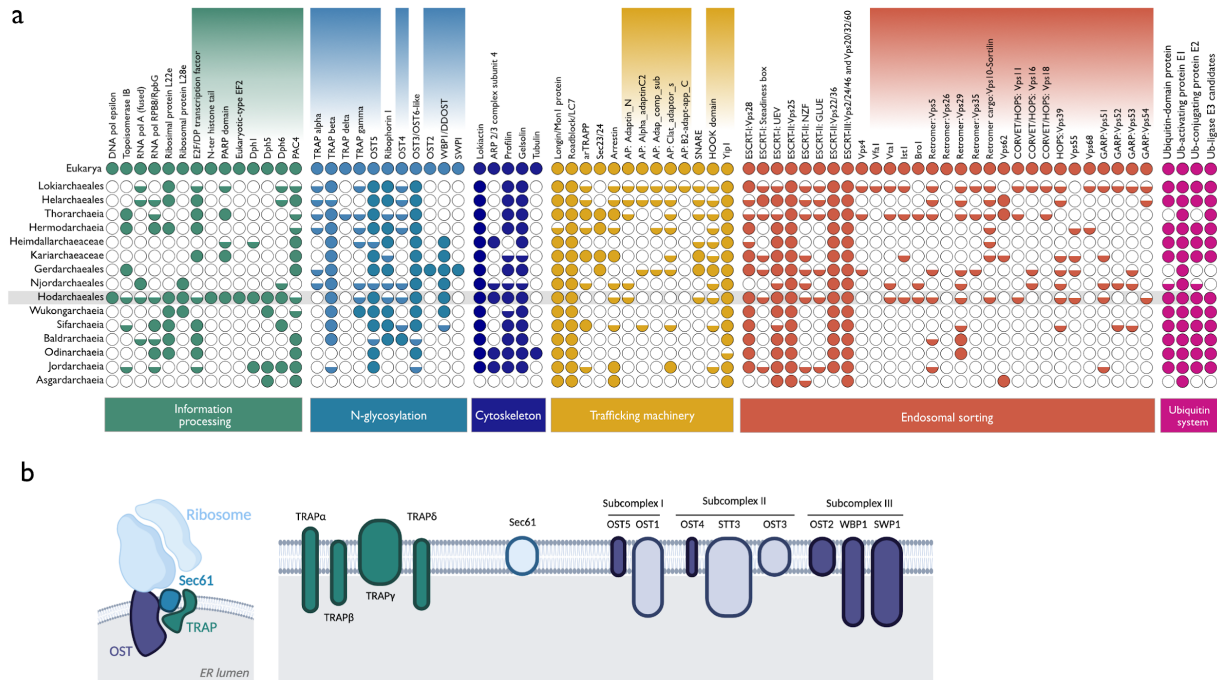
Figure 3. Eukaryotic signature proteins in Asgard archaea. a. Distribution of ESP homologs in Asgard archaea grouped by function. Shaded rectangles above protein names indicate ESPs newly identified as part of this study. Predicted homologs are depicted by colored circles: fully filled circles indicate that we detected homologs in at least half of the representative genomes of the clade; half-filled circles indicate that we detected homologs in fewer than half of the representative genomes of the clade. Hodarchaeales homologs are highlighted with a grey background. Accession numbers are available in Supplementary Table 3. b. Asgard archaea encode homologs of eukaryotic protein complexes involved in N-glycosylation. The Sec61, the OST and TRAP complexes are depicted according to their eukaryotic composition and localization. On the right-hand side of the panel, dark-colored subunits represent eukaryotic proteins which have prokaryotic homologs in Asgard archaea newly identified as part of this work; Light-colored subunit homologs have been described previously[10]. Figure generated with Biorender.com.

34

749

750

751

Figure 4. Genome dynamics, Optimal Growth Temperature predictions, and metabolic reconstruction of Asgard ancestors. a. Duplication (upper panel) and loss rates (lower panel) inferred for Asgard archaeal ancestors, normalized by proteome size and plotted by phylum. P-values for each Wilcoxon test against the median values of internal nodes belonging to TACK and Euryarchaeota are shown above each category, where *: p-value <= 0.05, **: p-value <=0.01, ***: p-value <=0.001. b. Optimal Growth Temperature (OGT) predictions, in degrees Celsius. OGT were predicted for the genomes presented here based on genomic and proteomic features[56] (Supplementary Table 5). Since nucleotide fractions of the ribosomal RNAs are used in this method, only those genomes with predicted rRNA genes could be analyzed. The right-hand panel depicts OGT within Heimdallarchaeia. Note that Njord- and Gerdarchaeales are predicted to be thermophiles (most genomes encode a reverse gyrase). In contrast, Hodarchaeales display the

35

763   lowest OGT among Heimdallarchaeia. c. Based on the presence/absence of the thermophily-

764   diagnostic enzyme reverse gyrase and the following metabolic signatures in each of the ancestors,

765   we predict that the last Asgard common ancestor probably transitioned from a hyperthermophilic

766   fermentative lifestyle to a mesophilic mixotroph lifestyle. The LAsCA likely encoded

767   gluconeogenic pathways *via* the reverse EMP gluconeogenic pathway and *via* FBP

768   aldolase/phosphatase (FBP A/P). The major energy-conserving step in the early Asgard ancestors

769   could have been the ATP synthesis by fermentation of small organic molecules (i.e., acetate,

770   formate, formaldehyde). The reverse ribulose monophosphate pathway (rRuMP) was a key

771   pathway in the LAsCA for the generation of reducing power. The Wood-Ljungdahl pathway

772   (WLP) appeared only to be present in the LAsCA and was lost in the more recent ancestors (of

773   Heimdallarchaeia and Hodarchaeales) indicated here. The tricarboxylic acid (TCA) cycle is

774   predicted to be complete in all five investigated Asgard ancestors. The inferred presence of the

775   electron transport chain (ETC) components is shown for selected ancestors of major Asgard

776   archaea groups, with the Hodarchaeales common ancestor encoding the most complete set of ETC

777   subunits, and likely using nitrate as a terminal electron acceptor. Therefore, membrane-associated

778   ATP biosynthesis coupled to the oxidation of NADH and succinate and reduction of nitrate to

779   nitrite within the respiratory chain could have been present in the LAECA. Abbreviations: Q:

780   quinone; c: cupredoxin; FBP A/P: Fructose 1,6-bisphosphate aldolase/phosphatase; EMP:

781   Embden-Meyerhof-Parnas; OPPP: Oxidative pentose phosphate pathway; rRuMP: Reversed

782   ribulose monophosphate pathway; RHP: reductive hexulose-phosphate; RuBisCO: Ribulose-1,5-

783   bisphosphate carboxylase/oxygenase; PRK: phosphoribulokinase; AMP: adenosine

784   monophosphate salvage pathway. Details of copy numbers of key enzymes involved in central

785   carbon metabolism are found in Supplementary Table 4.

36

Methods

Sample collection, sequencing, assembly and binning

We sampled aquatic sediments from eleven geographically distant sites: Guaymas Basin (Mexico), Lau Basin (Eastern Lau Spreading Center and Valu Fa Ridge, south-west Pacific Ocean), Hydrate Ridge (offshore of Oregon, USA), Aarhus Bay (Denmark), Radiata Pool (New Zealand), Taketomi Island Vent (Japan), the White Oak River estuary (USA), and Tibet Plateau and Tengchong (China) (Supplementary Table 1).

a. Jordarchaeote JZB50, QC4B49, QZMA23B3, QZMA2B5, QZMA3B5

Hot spring sediment samples were collected from Tibet Plateau and Yunnan Province (China) in 2016. The microbial community compositions have been described and reported previously[63,64]. Samples were collected from the hot spring pools using a sterile iron spoon into 50 ml sterile plastic tubes, then transported to the lab on dry ice, and stored at -80ºC for DNA extraction. The genomic DNA of the sediment samples was extracted using FastDNA Spin Kit for Soil (MP Biomedicals, Irvine, CA) according to the manufacturer's instructions. The obtained genomic DNA was purified for library construction, and sequenced on an Illumina HiSeq2500 platform (2 X 150 bp). The raw reads were filtered to remove Illumina adapters, PhiX and other Illumina trace contaminants with BBTools v38.79, and low-quality bases and reads using Sickle (v1.33; https://github.com/najoshi/sickle). The filtered reads were assembled using metaSPAdes (v3.10.1) with a kmer set of "21, 33, 55, 77, 99, 127". The filtered reads were mapped to the corresponding assembled scaffolds using bowtie2 v2.3.5.1[65]. The coverage of a given scaffold was calculated using the command of "jgi_summarize_bam_contig_depths" in MetaBAT v2.12.1[66]. For each sample, scaffolds with a minimum length of 2.5 kbp were binned into genome bins using MetaBAT v2.12.1, with both tetranucleotide frequencies (TNF) and scaffold coverage information

37

809   considered. The clustering of scaffolds from the bins and the unbinned scaffolds was visualized

810   using ESOM with a minimum window length of 2.5 kbp and max window length of 5 kbp, as

811   previously described[67]. Misplaced scaffolds were removed from bins and unbinned scaffolds

812   whose segments were placed within the bin areas of ESOMs were added to the corresponding bins.

813   Scaffolds with a minimum length of 1 kbp were uploaded to ggKbase

814   (http://ggkbase.berkeley.edu/). The ESOM-curated bins were further evaluated based on

815   consistency of GC content, coverage and taxonomic information, and scaffolds identified with

816   abnormal information were removed. The ggKbase genome bins were curated individually to fix

817   local assembly errors using ra2.py [68].

818   *b. Njordarchaeote A173, A3132, M288 and Thorarchaeote A361, A381, A399*

819   Hydrothermal vent deposits were collected from the ABE (ABE 1, 176° 15.48'W, 21° 26.68'S,

820   2142 m; ABE 3, 176° 15.59'W, 21° 26.95'S, 2131 m) and Mariner (176° 36.07'W, 22° 10.81'S,

821   1914 m) vent fields along the Eastern Lau Spreading Center in April/May of 2015 during the

822   RR1507 Expedition on the RV Roger Revelle. Sample collection and processing were done as

823   previously described[69]. DNA was extracted from homogenized rock slurries using the DNeasy

824   PowerSoil kit (Qiagen) as per the manufacturer's instructions. Samples were prepared for

825   sequencing on the Illumina HiSeq 3000 using Nextera DNA Library Prep kits (Illumina), and

826   metagenomes (2x150 bp) were sequenced at the Oregon State University Center for Genome

827   Research and Computing. Trimmomatic[70] v.0.36 was used to trim low-quality regions and adapter

828   sequences from raw reads (parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10,

829   LEADING:20, SLIDINGWINDOW:4:20, MINLEN:50). Clean paired reads were then interleaved

830   using the khmer software package[71]. Interleaved and unpaired reads were assembled with

831   MEGAHIT v.1.1.1-2-g02102e1 (--k-min 31, --k-max 151, --k-step 20, --min-contig-len 1000)

38

832   [72,73]. Trimmed reads were mapped back to the contigs to determine read coverage using Bowtie 2

833   v.2.2.9 [65,74] and SAMtools v.1.3.1[75]. Binning was performed with MetaBAT v.0.32.4[66] using

834   tetranucleotide frequency and read coverage. Bin completion and contamination were estimated

835   with CheckM v.1.0.7[76].

836   *c. Lokiarchaeote ABR01, ABR02, ABR03, ABR04, ABR05, ABR06, ABR08, ABR11, ABR13,*

837   *ABR15, Thorarchaeote ABR09, ABR10 and Heimdallarchaeote ABR14, ABR16* MAGs

838   were obtained as previously described[31].

839   *d. Archaeon WORA1, WORB2, Heimdallarchaeote WORE3, Lokiarchaeote WORB4,*

840   *WORC5 and Thorarchaeote WORH6*

841   Sampling, DNA extraction, sequencing library preparation and sequencing methods were

842   previously described[77]. Published assemblies and raw reads for the samples WOR-1-36_30

843   (SAMN06268458; Gp0056175), WOR-1-52-54 (SAMN06268416; Gp0059784), WOR-3-24_28

844   (SAMN06268417; Gp0059785) were downloaded from JGI. Short reads were trimmed using

845   Trimmomatic[70] v0.33 (PE ILLUMINACLIP:2:30:10 SLIDINGWINDOW:4:15 MILEN:100).

846   Contigs shorter than 1000 bp were excluded from the assembly using SeqTK v1.0r75

847   (https://github.com/lh3/seqtk). Each assembly was binned using CONCOCT v0.4.1[78] and coverage

848   information from the three datasets, and Asgard bins were subsequently identified based on

849   phylogenies of concatenated ribosomal proteins[10]. Identified Asgard MAGs were used together

850   with publicly available Asgard genomes to recruit trimmed-reads originated from Asgard genomes

851   using CLARK v1.2.3 with the -m 0 option[79]. For each dataset, recruited Asgard reads were

852   independently assembled using SPAdes[80] and IDBA-UD[81] and further binned using CONCOCT,

853   using a minimum contig length of 1000 bp. Bins with higher completeness and lower

854   contamination values as predicted by miComplete v1.00[82] were selected and manually curated

39

855   using mmgenome v0.7.1[83,84] using the coverage information, paired-reads linkage, composition

856   and marker genes information. The samples and assembly method used for each final MAG were:

857   Archaeon WORA1 (WOR-1-52-54; spades), Archaeon WORB2 (WOR-1-52-54; IDBA-UD),

858   Heimdallarchaeote WORE3 (WOR-3-24_28; spades), Lokiarchaeote WORB4 and WORC5

859   (WOR-1-36_30; IDBA-UD), and Thorarchaeote WORH6 (WOR-1-36_30; spades).

860   *e. Jordarchaeote RPD1, RPF2 and Odinarchaeote RPA3*

861   Information about the location of the hot spring sediments from Radiata Pool, sampling and DNA

862   extraction procedures was previously reported[10]. Short paired-end Illumina reads were generated

863   and preprocessed using Scythe (https://github.com/vsbuffalo/scythe) and Sickle

864   (https://github.com/najoshi/sickle) to remove adapters and low-quality reads. Reads were

865   subsequently assembled of IDBA-UD 1.1.3 (--maxk 124). The Jordarchaeote RPF2 MAG was

866   generated by binning contigs according to their tetranucleotide frequencies using esomWrapper.pl

867   (https://github.com/tetramerFreqs/Binning) with a minimum contig length 5000 bp and a window

868   size of 10 Kbp. ESOM maps were manually delineated using the Databionic ESOM viewer

869   (http://databionic-esom.sourceforge.net/). Jordarchaeote RPD1 and Odinarchaeote RPA3 were

870   binned following the methodology described in above (section d), but re-assembling the recruited

871   reads only with IDBA-UD (--maxk 124)[81].

872   *f. Jordarchaeote GBS01, GBS02, GBS03, GBS04, GBS05, GBS06, GBS07,*

873   *Heimdallarchaeote GBS08, GBS09, GBS10, GBS11, Lokiarchaeote GBS14,*

874   *Njordarchaeote GBS15, GBS16, GBS17, GBS18, GBS19, GBS20, GBS21, GBS22, GBS23,*

875   *GBS24, GBS25, GBS26, TNS08 and Thorarchaeote GBS28, GBS29, GBS33, GBS34*

876   MAGs were obtained as described in [85].

877   *g. Heimdallarchaeote B3_JM_08* MAG was obtained as described in [86].

878   h. *Thorarchaeote OWC_bin2, OWC_bin3 and OWC_bin4* MAGs were obtained as described

879   in [31].

880   i. *Heimdallarchaeote GBS11*

881   Samples were made available by the Gulf Coast Repository (GCR) and were collected on the

882   Ocean drilling Program (ODP) Leg 204 at site 1244 (44°35.17N, 125°7.19W) on July 14th, 2002

883   (hole C and core 2). The ODP site is found at a water depth of 890 m on the eastern side of the

884   South Hydrate Ridge on the Cascadia Margin. This site has been well characterized physically and

885   geochemically[87]. Furthermore, the microbial community structure has been surveyed using 16S

886   rRNA gene sequencing[88,89]. Two sediment samples, designated DCO-2-5 (sample ID 1489929)

887   and DCO-2-7 (sample ID 1489924), were collected at a sediment depth of 12.40 and 14.96 m

888   below the seafloor, respectively, and stored at -80°C at GCR. A total amount of 10 g of each of

889   the two sediment samples was used to extract DNA using the MoBio DNA PowerSoil Total kit. A

890   total amount of 100 ng DNA was used to prepare sequencing libraries that were 150 bp paired-end

891   sequenced at the Marine Biological Laboratory (Woods Hole, MA, USA) on an Illumina MiSeq

892   sequencer. Adaptors and DNA spike-ins were removed from the forward and reverse reads using

893   cutadapt   v1.12[90].   Afterwards,   reads   were   interleaved   using   interleave_fasta.py

894   (https://github.com/jorvis/biocode/blob/master/fasta/interleave_fasta.py),   and   further   trimmed

895   using Sickle with default settings (Fass JN) (https://github.com/najoshi/sickle). Metagenomic

896   reads from both samples were co-assembled using IDBA-UD using the following parameters:

897   --pre_correction, -mink 75, -maxk 105, --step 10, --seed_kmer 55 [81]. Metagenomic binning was

898   performed on scaffolds with a length >3,000 bp using ESOM, including a total of 4,939 scaffolds

899   with a length of 30,693,002 bp[67,81]. CheckM v1.0.5 was employed to evaluate the accuracy of the

900   binning approach by determining the percentage of completeness and contamination[76].

41

901   *j.   Heimdallarchaeote GBS09*

902   MAG was obtained as previously described[91].

903

904   **Exploration of phylogenetic diversity in Asgard assemblies and MAGs**

905   To assess the presence of potential Asgard-related lineages in our assemblies, we reconstructed a

906   phylogeny of ribosomal proteins encoded in a conserved 15-ribosomal protein (RP15) gene

907   cluster[16]. As ingroup, we used all MAGs presented in this study, plus all genomes classified as

908   Asgard archaea in NCBI as of June 25th 2021, plus those classified as "archaeon" corresponding

909   to   Hermodarchaeia   (GCA_016550385.1,   GCA_016550395.1,   GCA_016550405.1,

910   GCA_016550415.1,   GCA_016550425.1,   GCA_016550485.1,   GCA_016550495.1,

911   GCA_016550505.1). and all Asgard archaeal MAGs released by Sun et al.[21]. To obtain an

912   adequate outgroup dataset, we downloaded all archaeal genomes from the Genome Taxonomy

913   Database[92], data revision 89, and selected one genome sequence per species-level cluster as

914   defined in https://data.gtdb.ecogenomic.org/releases/release89/89.0/sp_clusters_r89.tsv. We then

915   selected a set of 216 genomes classified as Bathyarchaeia, Nitrososphaeria and Thermoprotei, and

916   used them as outgroup. Genes were detected and individually aligned and trimmed as previously

917   described[10]. Ribosomal protein sequences were selected if they were encoded in a contig

918   containing at least five of the 15 ribosomal protein genes. ModelFinder[93] was run as implemented

919   in IQ-TREE v. 2.0-rc2 to identify the best model among all combinations of the LG, WAG, JTT

920   and Q.pfam models, as well as their corresponding mixture models by adding +C20, +C40 and

921   +C60, and the additional mixture models LG4M, LG4X, UL2 and UL3, with rate heterogeneity

922   (none, +R4 and +G4) and frequency parameters (none, +F). A PMSF approximation[94] of the

923   chosen model (WAG+C60+R4+F) was then used for a final reconstruction using 100 non-

42

924   parametric bootstrap pseudoreplicates for branch statistical support. The obtained tree revealed a

925   broad genomic diversity of Asgard lineages (Figure 1).

926

927   Environmental distribution of Asgard archaea

928   16S rRNA gene sequences were predicted with Barrnap v 0.9

929   (https://github.com/tseemann/barrnap) with the option "--kingdom arc". Since none of the two

930   Helarchaeales bins contained 16S rRNA gene sequences, helarchaeal 16S rRNA gene sequences

931   identified by Seitz et al.[50] were used as representatives of this phylum. These sequences were

932   submitted to the IMNGS platform as queries for Paraller similarity searches against all available

933   NCBI sequence read archive (SRA) samples with a 95% similarity threshold and a minimum

934   alignment size of 200 bp[95]. Available metadata for detected SRA samples were then manually

935   assessed to link environmental context descriptions for individual SRA samples to broader

936   environment categories. The sequence abundance output file generated by IMNGS was then

937   analysed using R and the package tidyverse to calculate the number of SRA samples belonging to

938   each environment per phylum[96].

939

940   Gene prediction

941   Gene prediction was performed using Prokka[97] v1.12 (prokka --kingdom Archaea --norrna --

942   notrna). rRNA genes and tRNA genes were predicted with Barrnap

943   (https://github.com/tseemann/barrnap) and tRNAscan-SE[96,98], respectively.

944

Optimal growth temperature prediction

946 Optimal growth temperatures were predicted for the genomes presented here based on genomic

947 and proteomic features[56] (see Supplementary Information). Since ribosomal RNAs nucleotide

948 composition are used in this method, only genomes with predicted rRNAs were analyzed.

949

Identification of homologous protein families

951 All-versus-all similarity searches of all predicted proteins from the A64 taxon selection (64

952 Asgard, 76 TACK, 43 Euryarchaeota and 41 DPANN archaea; see Supplementary Table 2) were

953 performed using diamond[99] blastp (--more-sensitive --evalue 0.0001 --max-target-seqs 0 --outfmt

954 6). The file generated was used to cluster protein sequences into homologous families using

955 SiLiX[100] v.1.2.10, followed by Hifix[101] v1.0.6. The identity and overlap parameters required by

956 Silix were set to 0.2 and 0.7, respectively, after inspecting a wide range of values (--ident [0.15,0.4]

957 and --overlap [0.55-0.9], with increments of 0.05) and selecting the values that maximized the

958 number of clusters containing at least 80% of the taxa.

959

Functional annotation of homologous protein families

961 Protein families, excluding singletons, were aligned using mafft-linsi[102] v7.402 and converted into

962 HHsearch format (.hhm) profiles using HHblits[103] v3.0.3. Profile-profile searches were

963 subsequently performed against a database containing profiles from EggNOG 4.5[104], arCOGs[105]

964 and PFAM databases[106] that had been previously converted to the hhm format using HHblits[103]

965 v3.0.3.

44

966

Automatic functional annotation of individual proteins

Individual proteins were annotated using the HMMscan tool of the HMMer suite against PFAM v32[106], Interproscan[107] 5.25-64.0, EggNOG mapper v0.12.7 against the NOG database v4.5[104], diamond aligner v0.9.9.110[99] against the *nr* database, arCOG[105], and GhostKoala annotation server[108]. Putative hydrogenases were further classified using HydDB[109].

972

Detailed analysis of ESPs

In-depth analysis of potential ESPs involved a combination of automatic screens and manual curation. We first manually searched for homologs of previously described ESPs[9,10,42] by using a variety of sequence similarity approaches such as BLAST, HMMer tools, profile-profile searches using HHblits, combined with phylogenetic inferences, and, in some cases, the Phyre2 structure homology search engine[103,110,111]. We did not use fixed cutoffs, as the e-value between homologs will vary depending on the protein investigated, hence the need for manual examination of potential homologs and a combination of lines of evidence.

In addition, to identify potential new ESPs, we first used our profile-profile searches against EggNOG and manually investigated Asgard orthologous groups which had a best hit to a eukaryotic-specific EggNOG cluster. We also extracted PFAM domains whose taxonomic distribution is exclusive to eukaryotes as per PFAM v32, and investigated cases where they represented the best domain hit in Asgard archaea sequences identified by HMMscan. Finally, we manually investigated dozens of proteins known to be involved in key eukaryotic functions based on our knowledge and literature searches. In Figure 2, we are only reporting cases based on the

988   strict cutoff that the diagnostic HMM profile had the best score among all profiles detected for a

989   protein. An exception was made for the ESCRT domain Vps28, Steadiness box, UEV, Vps25,

990   NZF, GLUE and Vps22 domains which are usually found in combination with other protein

991   domains and thus do not necessarily represent the best scoring domain in a protein even if they

992   represent true homologs.

993

994   **Phylogenetic analyses of concatenated proteins for species tree inference**

995   Two sets of phylogenetic markers were used to infer the species tree. The first one (RP56) is based

996   on a previously published dataset of 56 ribosomal proteins used to place the first assembled Asgard

997   genomes[10]. The second one (NM57, for 'new markers') corresponds to 57 proteins extracted from

998   a set of 200 markers previously identified as core archaeal proteins that can be used to robustly

999   infer the tree of archaea[112]. These 57 markers were selected because they were found in at least a

1000   third of representatives of each of the 11 Asgard clades, as well as in 10 out of 14 eukaryotes, and

1001   were inherited from archaea in eukaryotes.

1002   We initially assembled an RP56 dataset for a phylogenetically diverse set of 222 archaeal

1003   and 14 eukaryotic taxa. These included all 11 Asgard archaea MAGs and genomes available at the

1004   NCBI as of May 12, 2017, as well as the 53 most diverse novel MAGs from this work (out of 63).

1005   We gathered orthologs of these genes from all proteomes by using sequences from the previously

1006   published alignment[10,112] as queries for BLASTp. For each marker, the best BLAST hit from each

1007   proteome was added to the dataset. For the first iteration, each dataset was aligned using mafft-

1008   linsi[113] and ambiguously aligned positions were trimmed using BMGE (-m BLOSUM30)[114]. All

1009   56 trimmed RP alignments were concatenated into an RP56-A64 supermatrix (236 taxa including

46

1010   64 Asgard archaea, 6332 amino acid positions). Once this taxon set was gathered, we identified

1011   homologs of the NM57 gene set as described above, thus generating supermatrix NM57-A64 (236

1012   taxa, 14,847 amino acid positions).

1013   We carried out a large number of phylogenomic analyses on variations of these two RP56-

1014   A64 and NM57-A64 datasets with different phylogenetic algorithms. Notably, preparing these

1015   datasets must be done with great care and is therefore time-consuming, and subsequent

1016   phylogenomic analyses generally require an enormous amount of computational running time.

1017   However, the rapid expansion of available Asgard archaeal MAGs, notably by Liu and colleagues

1018   as of April 2021[15], urged us to update and re-run many of the computationally demanding analyses.

1019   As some of the work that was based on a more restrained taxon sampling is still deemed valuable,

1020   such as some of the Bayesian phylogenomic analyses and ancestral genome content

1021   reconstructions, we retained these in the present study.

1022   An updated Asgard archaeal genomic sequence dataset was constructed by including all

1023   230 Asgard archaeal MAGs and genomes available at the NCBI database as of May 12, 2021, as

1024   well as 63 novel MAGs described in the present work. All 56 trimmed RP alignments were

1025   concatenated into an RP56-A293 supermatrix (465 taxa including 293 Asgard archaea, 7112 amino

1026   acid positions), which was used to infer a preliminary phylogeny with FastTree v2[115]

1027   (Supplementary Figure 16). Given the high computational demands of the subsequent analyses,

1028   we then used this phylogeny to select a subsample of Asgard archaea representatives. For this, we

1029   first removed the most incomplete MAGs encoding fewer than 19 ribosomal proteins (i.e., 1/3 of

1030   the markers) in the matrix. We also used the preliminary phylogeny to sub-select among closely

1031   related taxa: among taxa that were separated by branch lengths of <0.1, we only kept one

1032   representative. This led to a selection of 331 genomes, including 175 Asgard archaea, 41 DPANN,

47

1033   43 Euryarchaeota, and 72 TACK representatives (RP56-A175 dataset). Out of these 175 Asgard

1034   archaea, 41 correspond to MAGs newly reported here. Once this taxon set was gathered, we

1035   identified homologs of the NM57 gene set as described above, thus generating supermatrix NM57-

1036   A175. All datasets and their composition are summarized in Supplementary Table 2.

1037   To test for potential phylogenetic reconstruction artefacts, our datasets were subjected to several

1038   treatments. Supermatrices were recoded into four categories, using the SR4 scheme[27]. The

1039   corresponding phylogenies were reconstructed with IQ-TREE (using a user-defined previously

1040   described model referred to as 'C60SR4', based on the implemented 'LG+C60' model and

1041   modified to analyze the recoded data[10]) and Phylobayes (under the CAT+GTR model)[10]. We also

1042   used the estimated site rate output generated by IQ-TREE (-wsr) to classify sites into 10 categories,

1043   from the fastest to the slowest evolving, and we removed them in a stepwise fashion, removing

1044   from 10% to 90% of the data. Finally, we combined both approaches by applying SR4 recoding to

1045   the alignments obtained after each fast-site removal step. All phylogenetic analyses performed are

1046   summarized in Supplementary Table 2. See Supplementary Information for details and discussion.

1047

1048   **Analyses of individual proteins**

1049   For individual proteins of interest, we gathered homologs using various approaches, depending on

1050   the level of conservation across taxa. In order to detect putative Asgard homologs of eukaryotic

1051   proteins, we used a combination of tools including BLASTp[116] and the HMMer toolkit

1052   (http://hmmer.org/) if HMM profiles were available, and queried a local database containing our

1053   240 archaeal representatives (including all Asgard predicted proteomes). We then investigated the

1054   Asgard candidates by 1) using them as seed for BLASTp searches against nr; 2) by 3D modelling

1055   using Phyre2 and Swissmodel when sequence similarity was low; 3) by annotating them using

1056   Interproscan 5.25-64.0[107], EggNOG mapper v0.12.7[117], against the NOG database[117], and

1057   GhostKoala annotation server[108]; 4) by annotating the archaeal orthologous cluster they belonged

1058   to using profile-profile annotation as described above. Eukaryotic homologs were gathered from

1059   the UniRef50 database[118]. Depending on the divergence between homologs, they were aligned

1060   using mafft-linsi and trimmed using TrimAl[119] (--automated1) or BMGE[114], or, in cases where we

1061   investigated a specific functional domain, we used the hmmalign tool from the HMMer package

1062   with the --trim flag to only keep and align the region corresponding to this domain. When

1063   divergence levels allowed, phylogenetic analyses were performed using IQ-TREE with model

1064   testing including the C-series mixture models (-mset option)[120]. Statistical support was evaluated

1065   using 1000 ultrafast bootstrap replicates (for IQ-TREE)[119].

1066

1067   **Ancestral reconstruction**

1068   For the ancestral reconstruction analyses, only a subset of 181 taxa were included (64 Asgard, 74

1069   TACK and 43 Euryarchaeota, see Supplementary Table 2 for details). Protein families with more

1070   than three members were aligned and trimmed using mafft-linsi v7.402[113] and trimAl v1.4.rev15

1071   with the --gappyout option[119]. Tree distributions for individual protein families were estimated

1072   using IQ-TREE v1.6.5 (-bb 1000 -bnni -m TESTNEW -mset LG -madd LG+C10,LG+C20 -seed

1073   12345 -wbtl -keep-ident)[122]. The species phylogeny together with the gene tree distributions were

1074   subsequently used to compute 100 gene-tree species tree reconciliations using ALEobserve v0.4

1075   and ALEml_undated[52,53], including the fraction_missing option that accounts for incomplete

1076   genomes. The genome copy number was corrected to account for the extinction probability per

1077   cluster (github.com/maxemil/ALE/commit/136b78e). The missing fraction of the genome was

1078   calculated as 1 minus the completeness values (in fraction) as estimated by CheckM v1.0.5 for

1079   each of the 181 taxa[76]. Protein families containing only one protein (singletons) were considered

1080   as originations at the corresponding leaf. The ancestral reconstruction of 5 protein families that

1081   included more than 2000 proteins raised errors and could not be computed. The minimum

1082   threshold of the raw reconciliation frequencies for an event to be considered was set to 0.3 as

1083   commonly done[123–126] and recommended by the authors of ALE (Gergely Szölősi, personal

1084   communication).

1085

1086   **Ancestral metabolic inferences**

1087   Metabolic reconstruction of the Asgard ancestors was based on the inference, annotation and copy

1088   number of genes in ancestral nodes. The presence of a given gene was scored if its copy number

1089   in the ancestral nodes was above 0.3. A protein family was scored as "maybe present" if the

1090   inferred copy number was between 0.1 and 0.3. The protein annotation of each of the clusters

1091   containing the ancestral nodes was manually verified for each of the enzymatic steps involved in

1092   the pathways detailed in Supplementary Table 4.

1093

1094   # Data availability

1095   The MAGs reported in this study have been deposited at DDBJ/EMBL/GenBank. BioProject IDs,

1096   BioSample IDs and GenBank assembly accession numbers are available in Supplementary Table

1097   1 and will be released upon publication of the manuscript. All raw data underlying phylogenomic

1098   analyses (raw and processed alignments and corresponding phylogenetic trees) will be deposited

1099 on Figshare (https://figshare.com/account/home#/projects/111912) upon publication of the

1100 manuscript.

1101

Methods references

Acknowledgements

51

Author contributions

1140   T.J.G.E. conceived and supervised the study. A.S., K.U.K, W.H.L., Z.-S.H., A.L.R., W.-J.L., T.N.,

1141   A.-L.R., M.B.S., and A.P.T. collected and provided environmental samples. E.F.C., F.H., J.H.S.,

1142   N.D., K.W.S., B.J.B., L-X.C., J.F.B., and E.St.J. performed metagenomic sequence assemblies

1143   and metagenomic binning analyses. L.E., D.T., E.F.C., K.W.S., C.W.S., J.L., B.J.B., and T.J.G.E.

1144   analyzed genomic data. L.E., D.T., E.F.C. and F.H. performed phylogenomic analyses. L.E., D.T.,

1145   E.F.C., C.W.S., J.L. and T.J.G.E. investigated ESPs. E.F.C., L.E., and M.E.S. performed ancestral

1146   genome reconstruction analyses. V.D.A., B.J.B., C.W.S, L.E. and T.J.G.E. carried out metabolic

1147   inferences. L.E., D.T., E.F.C., C.W.S., V.D.A, B.J.B. and T.J.G.E. wrote, and all authors edited

1148   and approved, the manuscript.

1149   These authors contributed equally: Laura Eme, Daniel Tamarit, Eva F. Caceres.
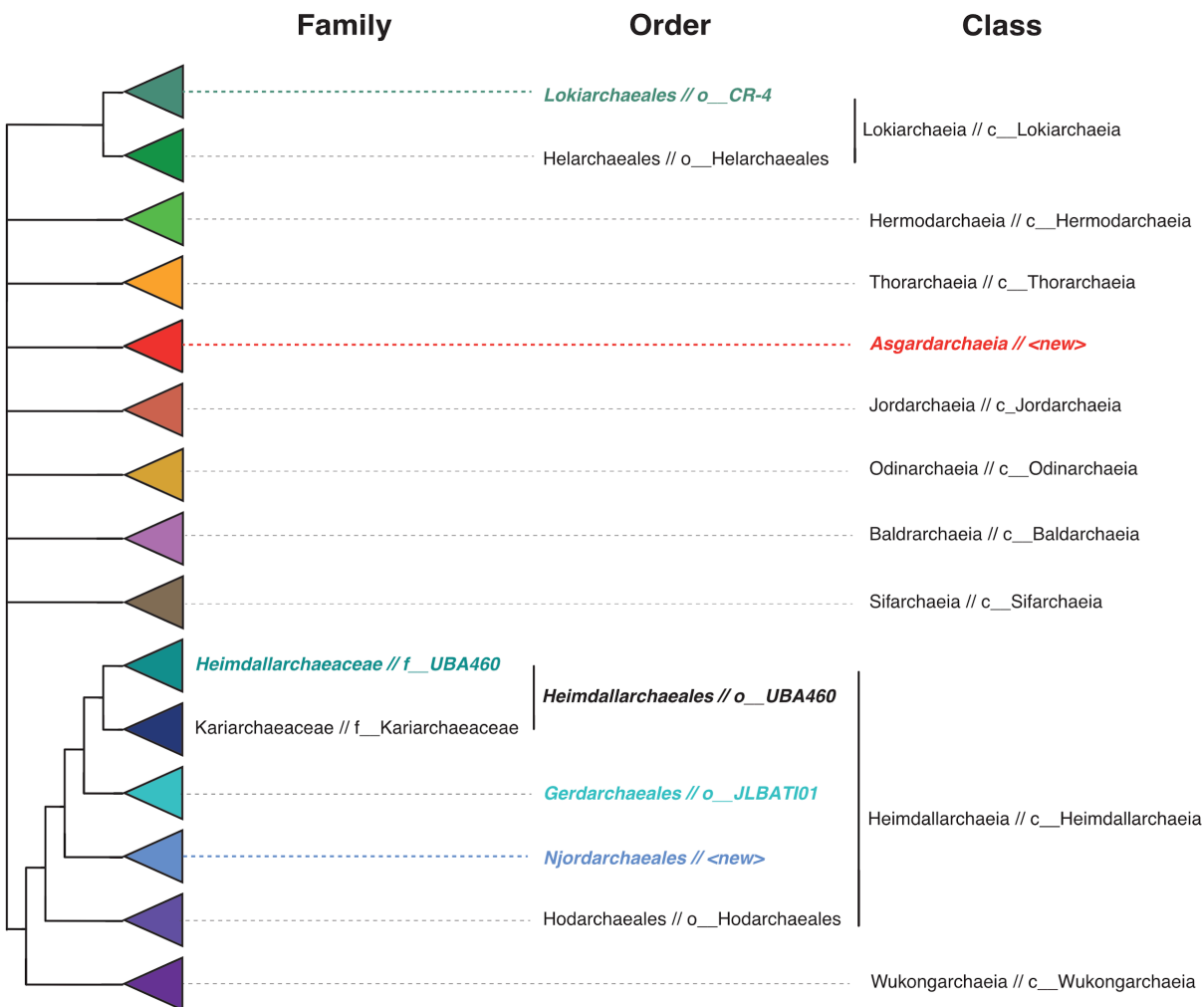
1150

Competing interest declaration

1151

1152   The authors declare no competing financial interests.

1153

1154

Extended data figures

Extended Data Figure 1. Cladogram of proposed taxonomic scheme for the ranks of family, order and class for Asgard archaeal lineages employed in this study. Equivalent names in GTDB are shown after a double slash (//). Cases with differing or new names have been highlighted in colored bold italics.

1163
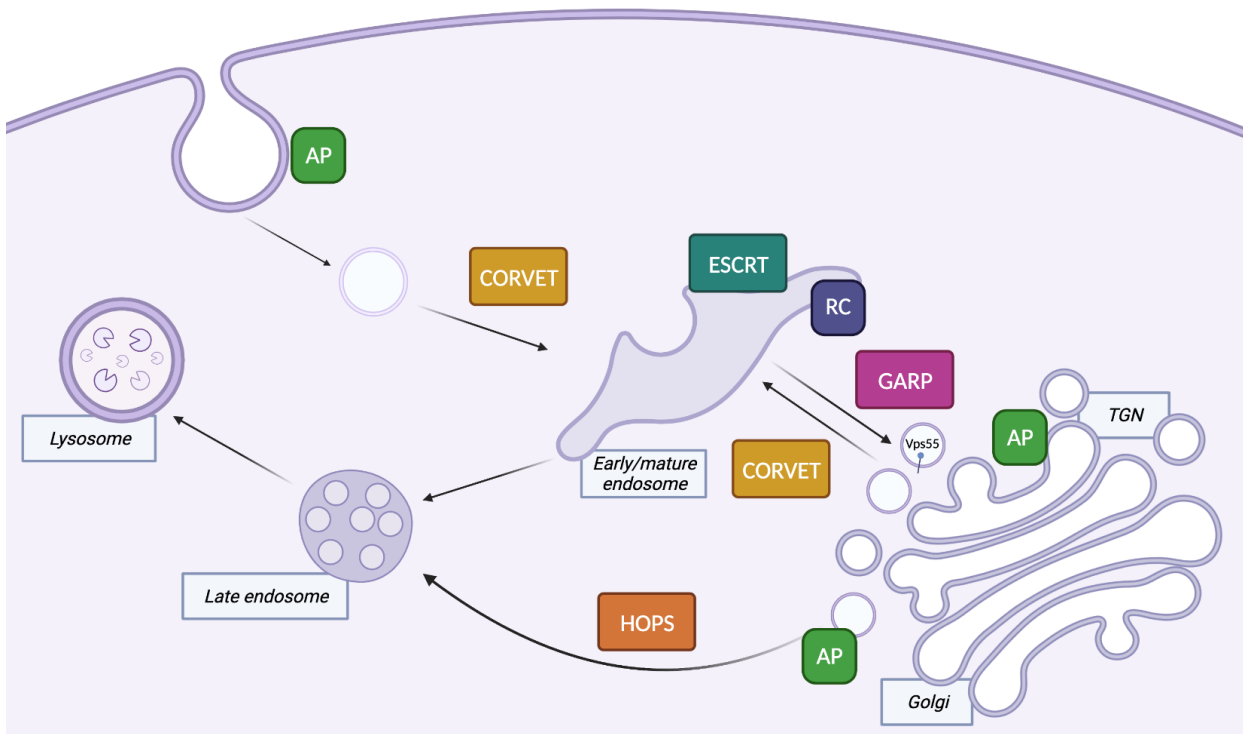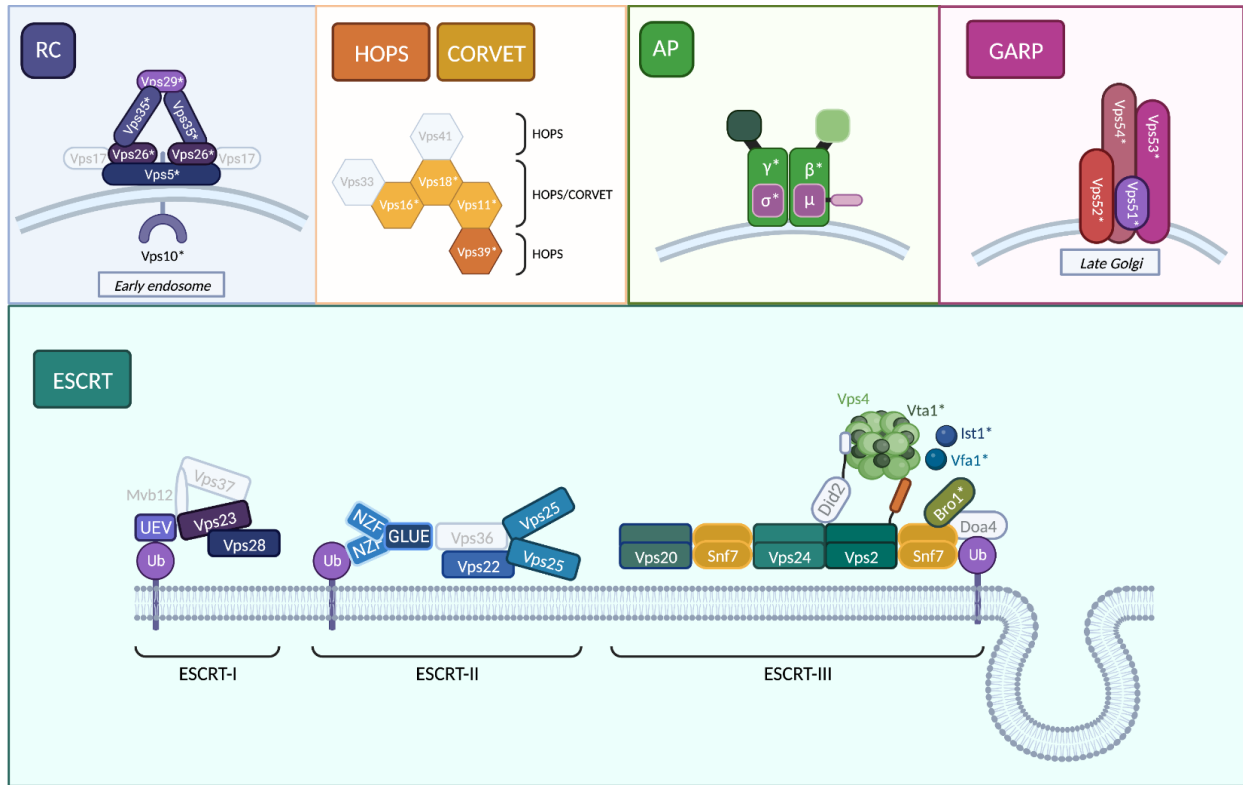
1164

1165

55

1166 **Extended Data Figure 2. Identification of previously undetected vesicular trafficking ESPs**

1167 **in Asgard archaea.** Schematic representation of a eukaryotic cell in which ESPs involved in

1168 membrane trafficking and endosomal sorting that have been identified in Asgard archaea are

1169 highlighted. Colored subunits have been detected in some Asgard archaea while grey ones seem

1170 to be absent from all current representatives. Only major protein complexes are depicted.

1171 Additional components can be found in Figure 2. From left to right, top to bottom: RC, Retromer

1172 complex. Retromer is a coat-like complex associated with endosome-to-Golgi retrograde traffic[36].

1173 It is formed by Vacuolar protein sorting-associated protein 35, Vps5, Vps17, Vps26 and Vps29[127].

1174 During cargo recycling, retromer is recruited to the endosomal membrane via the Vps5-Vps17

1175 dimer. Cargo recognition is thought to be mediated primarily through Vps26 and possibly by

1176 Vps35. Finally, the BAR domains of Vps5-Vps17 deform the endosomal membrane to form cargo-

1177 containing recycling vesicles. Their distribution is sparse, but we have detected Asgard archaeal

1178 homologs of all subunits except for Vps17. Interestingly, the Thorarchaeota Vps5-BAR domain is

1179 often fused to Vps28, a subunit of the ESCRT machinery complex I, suggesting a functional link

1180 between BAR domain proteins and the thorarchaeial ESCRT complex. The best-characterized

1181 retromer cargo is Vps10. This transmembrane protein receptor is known in yeast and mammal

1182 cells to be involved in the sorting and transport of lipoproteins between the Golgi and the

1183 endosome. The Vps10 receptor releases its cargo to the endosome and is recycled back to the Golgi

1184 via the retromer complex[128]. CORVET: Class C core vacuole/endosome tethering complex;

1185 HOPS: Homotypic fusion and protein sorting complex. Endosomal fusion and autophagy depend

1186 on the CORVET and HOPS hexameric complexes[39]; they share the core subunits Vps11, Vps16,

1187 Vps18, and Vps33[40]. In addition, HOPS is composed of Vps41 and Vps39[41]. Vps39, found

1188 associated to late endosomes and lysosomes, promotes endosomes/lysosomes clustering and their

1189 fusion with autophagosomes[129]. AP, Adaptor Proteins. Asgard archaea genomes from diverse

1190 phyla encode key functional domains of the AP complexes. The eukaryotic AP tetraheteromeric

1191 structure is depicted, each color corresponding to a PFAM functional domain (Medium green:

1192 Adaptin, N terminal region; Dark green: Alpha adaptin, C-terminal domain; Light green: Beta2-

1193 adaptin appendage, C-terminal sub-domain; Dark pink/clear outline: Clathrin adaptor complex

1194 small chain; Light pink/dark outline: C-ter domain of the mu subunit); all five domains were

1195 detected in Asgard archaea, although not fused to each other. GARP: Golgi-associated retrograde

1196 protein complex. The GARP complex is a multisubunit tethering complex located at the trans-

1197 Golgi network where it functions to tether retrograde transport vesicles derived from

1198 endosomes[37,38]. GARP comprises four subunits, VPS51, VPS52, VPS53, and VPS54. ESCRT:

1199 Endosomal Sorting Complex Required for Transport system. This complex machinery performs a

1200 topologically unique membrane bending and scission reaction away from the cytoplasm. While

1201 numerous components of the ESCRT-I, II and III systems have been previously detected in Asgard

1202 archaea[9,10,42], we here report Asgard homologs for several ESCRT-III regulators Vfa1, Vta1, Ist1,

1203 and Bro1. The bottom panel shows where these complexes mainly act in eukaryotic cells. Ub:

1204 Ubiquitin; Vps: vacuolar protein sorting. Subunit names in grey indicate that no homologs were

1205 detected in Asgard archaea. Domains newly identified as part of this study are indicated with an

1206 asterisk. Created with BioRender.com.

1207

1208

Supplementary information

1209

1210

1211 **Supplementary Information.** This file contains Supplementary Methods, Supplementary

1212 Discussions, Supplementary Figures 1-32, Supplementary Tables 1-8, Supplementary Data and

1213 Supplementary References.

1214

1215 Correspondence and requests for materials should be addressed to thijs.ettema@wur.nl.