nature portfolio

https://doi.org/10.1038/s41586-021-03494-3

Supplementary information

Expanded diversity of Asgard archaea and their relationships with eukaryotes

In the format provided by the authors and unedited

Expanded diversity of Asgard archaea and their relationships with eukaryotes

Yang Liu[1,5], Kira S. Makarova[2,5], Wen-Cong Huang[1,5], Yuri I. Wolf[2], Anastasia Nikolskaya[2], Xinxu Zhang[1], Mingwei Cai[1], Cui-Jing Zhang[1], Wei Xu[3], Zhuhua Luo[3], Lei Cheng[4], Eugene V. Koonin[2]† & Meng Li[1]†

[1]Shenzhen Key Laboratory of Marine Microbiome Engineering, Institute for Advanced Study, Shenzhen University, Shenzhen, Guangdong, 518060, P. R. China

[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

[3]Key Laboratory of Marine Biogenetic Resources, Third Institute of Oceanography, Ministry of Natural Resources, Xiamen, P. R. China

[4]Key Laboratory of Development and Application of Rural Renewable Energy, Biogas Institute of Ministry of Agriculture, Chengdu 610041, P.R. China

[5]These authors contributed equally: Yang Liu, Kira S. Makarova, Wen-Cong Huang.

†e-mail: koonin@ncbi.nlm.nih.gov; limeng848@szu.edu.cn

1. Description of the new taxa

'*Candidatus* Wukongarchaeum' (Wu.kong.ar.chae'um. N. L. n. Wukong a legendary Chinese figure, also known as Monkey King, who caused havoc in the heavenly palace); N.L. neut. N. *archaeum* (from Gr. adj. archaios ancient) archaeon; N. L. neut. N. Wukongarchaeum.

'*Candidatus* Wukongarchaeum yapensis' (yap'ensis N. L. masc. adj. pertaining to Yap trench, which is the geographical position where the first type material of this species was obtained). Type material is the genome designated as As_085 (Yap4.bin4.70) representing '*Candidatus* Wukongarchaeum yapensis'. The genome "As_085" represents a MAG consisting of 2.16 Mbps in 277 contigs with an estimated completeness of 92.52%, an estimated contamination of 4.05%, a 16S and 23S rRNA gene and 14 tRNAs. The MAG recovered from a marine water metagenome (Yap trench, Western Pacific), with an estimated depth of coverage of 31.4, has a GC content of 38%.

***Candidatus* Wukongarchaeaceae** (Wu.kong.ar.chae.a.ce'ae. N.L. neut. n. Wukongarchaeum a (Candidatus) type genus of the family; -aceae ending to denote the family; N.L. fem. pl. n. Wukongarchaeaceae the Wukongarchaeum family).

The family is delineated based on 209 concatenated Asgard Cluster of Orthologs (AsCOGs) and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Wukongarchaeum.

***Candidatus* Wukongarchaeales** (Wu.kong.ar.chae.a'les. N.L. neut. n. Wukongarchaeum a (Candidatus) type genus of the order; -ales ending to denote the order; N.L fem. pl. n. Wukongarchaeales the Wukongarchaeum order).

The order is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is Candidatus Wukongarchaeum.

***Candidatus* Wukongarchaeia** (Wu.kong.ar.chae'i.a. N.L. neut. n. Wukongarchaeum a (Candidatus) type genus of the order of the class; -ia ending to denote the class; N.L fem. pl. n. Wukongarchaeia the Wukongarchaeum class).

The class is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type order *Candidatus* Wukongarchaeales.

***Candidatus* Wukongarchaeota** (Wu.kong.ar.chae.o'ta. N.L. neut. n. Wukongarchaeum a (Candidatus) type genus of the class of the phylum; -ota ending to denote the phylum; N.L neut. pl. n. Wukongarchaeota the Wukongarchaeum phylum)

The phylum is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type class *Candidatus* Wukongarchaeia.

'*Candidatus* Hodarchaeum' (Hod.ar.chae'um. N. L. n. Hod a son of Odin in Norse mythology); N.L. neut. N. archaeum (from Gr. adj. archaios ancient) archaeon; N. L. neut. N. Hodarchaeum.

'*Candidatus* Hodarchaeum mangrovi' (man.gro'vi N.L. fem. n. of a mangrove, referring to the isolation of the type material from mangrove soil). Type material is the genome designated as As_027

(FT2_5_011) representing '*Candidatus* Hodarchaeum mangrovi'. The genome "As_027" represents a MAG consisting of 4.01 Mbps in 348 contigs with an estimated completeness of 93.61%, an estimated contamination of 0.93%, a 23S rRNA gene and 14 tRNAs. The MAG recovered from mangrove sediment metagenomes (Futian Nature Reserve, China), with an estimated depth of coverage of 17.9, has a GC content of 32.9%.

***Candidatus* Hodarchaeaceae** (Hod.ar.chae.a.ce'ae. N.L. neut. n. Hodarchaeum a (Candidatus) type genus of the family; -aceae ending to denote the family; N.L. fem. pl. n. Hodarchaeaceae the Hodarchaeum family).

The family is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Hodarchaeum.

***Candidatus* Hodarchaeales** (Hod.ar.chae.a'les. N.L. neut. n. Hodarchaeum a (Candidatus) type genus of the order; -ales ending to denote the order; N.L fem. pl. n. Hodarchaeales the Hodarchaeum order).

The order is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Hodarchaeum.

***Candidatus* Hodarchaeia** (Hod.ar.chae'i.a. N.L. neut. n. Hodarchaeum a (Candidatus) type genus of the order of the class; -ia ending to denote the class; N.L fem. pl. n. Hodarchaeia the Hodarchaeum class).

The class is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type order *Candidatus* Hodarchaeales.

***Candidatus* Hodarchaeota** (Hod.ar.chae.o'ta. N.L. neut. n. Hodarchaeum a (Candidatus) type genus of the class of the phylum; -ota ending to denote the phylum; N.L neut. pl. n. Hodarchaeota the Hodarchaeum phylum)

The phylum is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type class *Candidatus* Hodarchaeia.

'*Candidatus* Kariarchaeum' (Ka.ri.ar.chae'um. N. L. n. Kari the god of wind in Norse mythology); N.L. neut. N. *archaeum* (from Gr. adj. archaios ancient) archaeon; N. L. neut. N. Kariarchaeum.

'*Candidatus* Kariarchaeum pelagius' (pe.la'gi.us. L. masc. adj. of or belonging to the sea, referring to the isolation of the type material from the Ocean). Type material is the genome designated as As_030 (RS678) representing '*Candidatus* Kariarchaeum pelagius'. The genome "As_030" represents a MAG consisting of 1.41 Mbps in 76 contigs, an estimated completeness of 83.18%, with an estimated contamination of 1.87%, a 23S, 16S and 5S rRNA genes and 18 tRNAs. The MAG recovered from a marine metagenome (Saudi Arabia: Red Sea) has a GC content of 30.11%.

***Candidatus* Kariarchaeaceae** (Ka.ri.ar.chae.a.ce'ae. N.L. neut. n. Kariarchaeum a (Candidatus) type genus of the family; -aceae ending to denote the family; N.L. fem. pl. n. Kariarchaeaceae the Kariarchaeum family).

The family is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Kariarchaeum.

***Candidatus* Kariarchaeales** (Ka.ri.ar.chae.a'les. N.L. neut. n. Kariarchaeum a (Candidatus) type genus of the order; -ales ending to denote the order; N.L fem. pl. n. Kariarchaeales the Kariarchaeum order).

The order is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Kariarchaeum.

***Candidatus* Kariarchaeia** (Ka.ri.ar.chae'i.a. N.L. neut. n. Kariarchaeum a (Candidatus) type genus of the order of the class; -ia ending to denote the class; N.L fem. pl. n. Kariarchaeia the Kariarchaeum class).

The class is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type order *Candidatus* Kariarchaeales.

***Candidatus* Kariarchaeota** (Ka.ri.ar.chae.o'ta. N.L. neut. n. Kariarchaeum a (Candidatus) type genus of the class of the phylum; -ota ending to denote the phylum; N.L neut. pl. n. Kariarchaeota the Kariarchaeum phylum)

The phylum is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type class *Candidatus* Kariarchaeia.

'*Candidatus* Borrarchaeum' (Borr.ar.chae'um. N. L. n. Borr a creator god and father of Odin); N.L. neut. N. *archaeum* (from Gr. adj. archaios ancient) archaeon; N. L. neut. N. Borrarchaeum.

'*Candidatus* Borrarchaeum yapensis' (yap'ensis N. L. masc. adj. pertaining to Yap trench, which is the geographical position where the first type material of this species was obtained). Type material is the genome designated as As_181 (Yap2000.bin9.141) representing '*Candidatus* Borrarchaeum yapensis'. The genome "As_181" represents a MAG consisting of 3.63 Mbps in 125 contigs, with an estimated completeness of 95.02%, an estimated contamination of 5.61% and 11 tRNAs. The MAG, recovered from a marine water metagenome (Yap trench, Western Pacific) with an estimated depth coverage of 15.04, has a GC content of 37.1%.

***Candidatus* Borrarchaeaceae** (Borr.ar.chae.a.ce'ae. N.L. neut. n. Borrarchaeum a (Candidatus) type genus of the family; -aceae ending to denote the family; N.L. fem. pl. n. Borrarchaeaceae the Borrarchaeum family).

The family is delineated based on 209 concatenated AsCOGs phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Borrarchaeum.

***Candidatus* Borrarchaeales** (Borr.ar.chae.a'les. N.L. neut. n. Borrarchaeum a (Candidatus) type genus of the order; -ales ending to denote the order; N.L fem. pl. n. Borrarchaeales the Borrarchaeum order).

The order is delineated based on 209 concatenated AsCOGs phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Borrarchaeum.

***Candidatus* Borrarchaeia** (Borr.ar.chae'i.a. N.L. neut. n. Borrarchaeum a (Candidatus) type genus of the order of the class; -ia ending to denote the class; N.L fem. pl. n. Borrarchaeia the Borrarchaeum class).

The class is delineated based on 209 concatenated AsCOGs phylogeny. The description is the same as that of its sole and type order *Candidatus* Borrarchaeales.

***Candidatus* Borrarchaeota** (Borr.ar.chae.o'ta. N.L. neut. n. Borrarchaeum a (Candidatus) type genus of the class of the phylum; -ota ending to denote the phylum; N.L neut. pl. n. Borrarchaeota the Borrarchaeum phylum)

The phylum is delineated based on 209 concatenated AsCOGs phylogeny. The description is the same as that of its sole and type class *Candidatus* Borrarchaeia.

'*Candidatus* Baldrarchaeum' (Bal.dr.ar.chae'um. N. L. n. Baldr the god of light and son of Odin and borther of Thor in Norse mythology); N.L. neut. N. *archaeum* (from Gr. adj. archaios ancient) archaeon; N. L. neut. N. Baldrarchaeum.

'*Candidatus* Baldrarchaeum yapensis' (yap'ensis N. L. masc. adj. pertaining to Yap trench, which is the geographical position where the first type material of this species was obtained). Type material is the genome designated as As_130 (Yap30.bin9.72) representing '*Candidatus* Baldrarchaeum yapensis'. The genome "As_130" represents a MAG consisting of 2.27 Mbps in 100 contigs, with an estimated completeness of 93.93%, an estimated contamination of 3.74%, a 23S and 16S rRNA gene and 15 tRNAs. The MAG, recovered from a marine water metagenome (Yap trench, Western Pacific) with an estimated depth coverage of 39.99, has a GC content of 45.95%.

***Candidatus* Baldrarchaeaceae** (Bal.dr.ar.chae.a.ce'ae. N.L. neut. n. Baldrarchaeum a (Candidatus) type genus of the family; -aceae ending to denote the family; N.L. fem. pl. n. Baldrarchaeaceae the Baldrarchaeum family).

The family is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Baldrarchaeum.

***Candidatus* Baldrarchaeales** (Bal.dr.ar.chae.a'les. N.L. neut. n. Bladrarchaeum a (Candidatus) type genus of the order; -ales ending to denote the order; N.L fem. pl. n. Baldrarchaeales the Baldrarchaeum order).

The order is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Baldrarchaeum.

***Candidatus* Baldrarchaeia** (Bal.dr.ar.chae'i.a. N.L. neut. n. Baldrarchaeum a (Candidatus) type genus of the order of the class; -ia ending to denote the class; N.L fem. pl. n. Baldrarchaeia the Baldrarchaeum class).

The class is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type order *Candidatus* Baldrarchaeales.

***Candidatus* Baldrarchaeota** (Bal.dr.ar.chae.o'ta. N.L. neut. n. Baldrarchaeum a (Candidatus) type genus of the class of the phylum; -ota ending to denote the phylum; N.L neut. pl. n. Baldrarchaeota the Baldrarchaeum phylum)

The phylum is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type class *Candidatus* Baldrarchaeia.

'*Candidatus* Hermodarchaeum' (Her.mod.ar.chae'um. N. L. n. Hermod, messengers of the gods in the Norse mythology and son of Odin and brother of Baldr in the Norse mythology); N.L. neut. N. *archaeum* (from Gr. adj. archaios ancient) archaeon; N. L. neut. N. Hermodarchaeum.

'*Candidatus* Hermodarchaeum yapensis' (yap'ensis N. L. masc. adj. pertaining to Yap trench, which is the geographical position where the first type material of this species was obtained). Type material is the genome designated as As_086 (Yap4.bin9.105) representing '*Candidatus* Hermodarchaeum yapensis'. The genome 'As_086' represent a MAG consisting of 2.71 Mbps in 77 contigs, with an estimated completeness of 92.99%, an estimated contamination of 1.87%, a 23S and 16S rRNA gene and 16 tRNAs. The MAG, recovered from a marine water metagenome (Yap trench, Western Pacific) with an estimated depth coverage of 19.24, has a GC content of 44.69%.

***Candidatus* Hermodarchaeaceae** (Her.mod.ar.chae.a.ce'ae. N.L. neut. n. Hermodarchaeum a (Candidatus) type genus of the family; -aceae ending to denote the family; N.L. fem. pl. n. Hermodarchaeaceae the Hermodarchaeum family).

The family is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Hermodarchaeum.

***Candidatus* Hermodarchaeales** (Her.mod.ar.chae.a'les. N.L. neut. n. Hermodarchaeum a (Candidatus) type genus of the order; -ales ending to denote the order; N.L fem. pl. n. Hermodarchaeales the Hermodarchaeum order).

The order is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole genus and species. Type genus is *Candidatus* Hermodarchaeum.

***Candidatus* Hermodarchaeia** (Her.mod.ar.chae'i.a. N.L. neut. n. Hermodarchaeum a (Candidatus) type genus of the order of the class; -ia ending to denote the class; N.L fem. pl. n. Hermodarchaeia the Hermodarchaeum class).

The class is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type order *Candidatus* Hermodarchaeales.

***Candidatus* Hermodarchaeota** (Her.mod.ar.chae.o'ta. N.L. neut. n. Hermodarchaeum a (Candidatus) type genus of the class of the phylum; -ota ending to denote the phylum; N.L neut. pl. n. Hermodarchaeota the Hermodarchaeum phylum)

The phylum is delineated based on 209 concatenated AsCOGs and 16S rRNA gene phylogeny. The description is the same as that of its sole and type class *Candidatus* Hermodarchaeia.

2. Clusters of orthologous genes of Asgard archaea

The previous analyses of Asgard genomes detected a large fraction of "dark matter" genes [69]. For example, in the recently published complete genome of *Candidatus* P. syntrophicum, 45% of the proteins are annotated as "hypothetical". Here, an effort was made to improve the annotation of Asgard genomes

194 by investigating this dark matter in greater depth and developing a dedicated platform for Asgard
195 comparative genomics. To this end, Asgard Clusters of Orthologous Genes (asCOGs) were constructed,
196 and the most sensitive available methods of sequence analysis were employed to annotate additional
197 Asgard proteins, attempting, in particular, to expand the catalogue of Asgard homologs of ESPs (see
198 Methods for details).

199 Preliminary clustering by sequence similarity and analysis of the protein cluster representation across the
200 genomes identified the set of 76 most complete Asgard MAGs (46 genomes available previously and 30
201 ones reported here) that cover most of the group diversity (Supplementary Table 1). The first version of
202 the asCOGs presented here consists of 14,704 orthologous protein families built for this 76-genome set.
203 The asCOGs cover from 72% to 98% (92% on average) of the proteins in these 76 genomes
204 (Supplementary Data file 1). Many asCOGs include individual domains of large, multidomain proteins.

205 The gene commonality plot for the asCOGs shows an abrupt drop at the right end, which reflects a
206 surprising deficit of nearly universal genes (Extended Data Fig. 3). Such shape of the gene commonality
207 curve appears anomalous compared to other major groups of archaea or bacteria with many sequenced
208 genomes [70]. For example, in the case of the TACK superphylum of archaea, for which the number of
209 genomes available is similar to that for Asgard, with a comparable level of diversity, the commonality
210 plot shows no drop at the right end, but instead, presents a clear uptick, which corresponds to the core of
211 genes represented in (almost) all genomes (Extended Data Fig. 3). Apparently, most of the Asgard
212 genomes remain incomplete, such that conserved genes were missed randomly. Currently, there are only
213 three gene families that are present in all Asgard MAGs, namely, a Zn-ribbon domain, a Threonyl-tRNA
214 synthetase and an aminotransferase (Supplementary Data File 1).

215 The asCOG profiles were employed to annotate the remaining 86 Asgard MAGs, including those that
216 were sequenced in the later stages of this work (Supplementary Table 1). On average, 89% of the proteins
217 encoded in these genomes were covered by asCOGs (Supplementary Table 1). The protein annotation
218 obtained using asCOGs was compared with the available annotation of '*Candidatus* P. syntrophicum'.
219 Using asCOGs allowed at least a general functional prediction for 649 of the 1756 (37%) 'hypothetical
220 proteins' in this organism, the only one in Asgard with a closed genome. We also identified 139 proteins,
221 in addition to the 80 described originally, that can be considered Eukaryotic Signature Proteins, or ESPs
222 (see below). Thus, the asCOGs database appears to be an efficient tool for annotation and comparative
223 genomic analysis of Asgard MAGs and complete genomes.

224

3. The core gene set of Asgard archaea

The core set of conserved Asgard genes was arbitrarily defined as all asCOGs that are present in at least one-third of the MAGs in each of the 12 phylum-level lineages, with the mean representation across lineages >75%. Under these criteria, the Asgard core includes 378 asCOGs (Extended Data Fig. 6b , Supplementary Table 6). As expected, most of these protein families, 293 (77%), are universal (present in bacteria, other archaea and eukaryotes), 62 (16%) are represented in other archaea and eukaryotes, but not in bacteria, 15 (4%) are found in other archaea and bacteria, but not in eukaryotes, 7 (2%) are archaea-specific, and only 1 (0.003%) is shared exclusively with eukaryotes (Extended Data Fig. 6c). Most of the core asCOGs show comparable levels of similarity to homologs from two or all three domains of life. The second largest fraction of the core asCOGs shows substantially greater sequence similarity (at least, 25% higher similarity score) to homologous proteins from archaea than to those from eukaryotes and/or bacteria (Supplementary Table 6). Compared with the 219 genes that comprise the pan-archaeal core [71], the Asgard core set lacks 12 genes, each of which, however, is present in some subset of the Asgard genomes. These include three genes of diphthamide biosynthesis and 2 ribosomal proteins, L40E and L37E. The intricate evolutionary history of gene encoding translation elongation factors and enzymes of diphthamide biosynthesis in Asgard has been analyzed previously [72]. Also of note is the displacement of the typical archaeal glyceraldehyde-3-phosphate dehydrogenase (type II) by a bacterial one (type I) in most of the Asgard genomes (cog.001204, Supplementary Data File 1).

Functional distribution of the core asCOGs is shown in Extended Data Fig. 6b (also see Supplementary Data File 1). For comparison, we also derived an extended gene core for the TACK superphylum, using similar criteria (at least 50% in each of the 6 lineages and 75% of the genomes overall, Extended Data Fig. 6b). For at least half of the Asgard core genes, across most functional classes, there were no orthologs in the TACK core. The most pronounced differences were found, as expected, in the category U (intracellular trafficking, secretion, and vesicular transport). In Asgard archaea, this category includes 19 core genes compared with 7 genes in TACK; 13 of these genes are specific to the Asgard archaea and include components of ESCRT I and II, 3 distinct Roadblock/longin families, 2 distinct families of small GTPases, and a few other genes implicated in related processes (Supplementary Table 6).

4. Phylogenomic analysis of the Asgard superphylum and Asgard-eukaryote evolutionary relationship

To assess the robustness of the Asgard phylogeny, 100 random, independent samples of 209 core asCOGs were generated, where each was sampled with the probability of $1-e^{-1}$ (equivalent to making a bootstrap sample of the 209 asCOGs with each asCOG included once). The concatenated alignment of each subsampled set was then analyzed using IQ-tree and the automatically chosen evolutionary model. The support for the tree topology, derived from the full set of 209 core asCOGs, was estimated from the bipartitions in these 100 trees (Fig. 1a).

The analysis of the universal phylogeny aimed to make the species set for phylogenetic reconstruction as broadly representative as possible, while keeping its size manageable, to allow the use of powerful phylogenetic methods. The tree was constructed from alignments of 30 families of conserved orthologous proteins of 162 Asgard archaea, 286 other archaea, 98 bacteria and 72 eukaryotes (see Methods for details of the procedure including the selection of a representative species set and Supplementary Table 4). This set of universal genes, for which comparatively little HGT has been identified, was selected and employed for a classical reconstruction of the tree of life [19] as well as in many subsequent studies on deep phylogeny [20,21]. A concatenated alignment of 7411 positions was generated for these 30 protein families, after removing low information content positions (Supplementary Table 4). For the phylogenetic reconstruction, we used the IQ-tree program with several phylogenetic models (see Methods and Supplementary Table 5 for details). The resulting tree had the 3D topology, with high support values for all key bifurcations (Extended Data Fig. 7a and Supplementary Data File 2).

The effect of the phylogenetic marker choice on the tree topology was investigated by: 1) generating 30 sets of 29 markers by removing each of the 30 markers from the original set and 2) generating 100 bootstrap-like random, independent samples of markers (in the same manner as with the Asgard phylogeny markers). The concatenated alignment of each subsampled set of markers was then analyzed using IQ-tree and the automatically chosen evolutionary model. Unexpectedly, the 3D topology completely hinged on the presence of a single marker, COG0012 (Ribosome-binding ATPase YchF, an essential protein involved in translation) [22]. All samples that included COG0012 strongly supported the 3D topology. In a sharp contrast, the set of 29 markers without COG0012 (Fig. 1c) and all 32 bootstrap-like samples that did not contain COG0012 equally strongly supported the 2D topology with varying placement of the eukaryote branch within archaea (Supplementary Table 5, Supplementary Data File 2). To further test the potential effect of branch-specific evolutionary rates on the tree topology, a 30-marker tree and a 29-marker tree without COG0012 were additionally contructed under a 4-category heterotachy model (LG+FO*H4) [73]. The resulting trees showed essentially the same topologies as the trees in Extended Data Fig. 7a and Fig. 1c, respectively (Supplementary Data file 2).

To further assess the effect of the choice of the evolutionary model and the species selection on the tree topology, 100 trees were constructed from the same alignment (29 markers, concatenated) by randomly sampling 5 representatives of Asgard archaea, other archaea, bacteria, and eukaryotes each (using the best model selected by IQ-tree). All 100 trees showed 2D topology with at least one non-Asgard archaea clade separating eukaryotes from bacteria (Supplementary Table 5, Supplementary Data file 2).

To analyze the compatibility of the phylogenetic signals from the individual marker alignments with that of the 29-marker tree built from the concatenated alignment, 1000 random five-species samples were generated from each of the 29 protein families by sampling one sequence from Asgard archaea, one from the TACK archaea, one from other archaea, one from eukaryotes and one from bacteria. The topology of the five-species species trees was assessed for compatibility with the topology of the corresponding species in the 29-marker tree. The probabilities, associated with the Approximately Unbiased (AU) test [74], implemented in the IQ-tree program, were averaged across the 1000 random samples. The trees for all 29 markers were found to be compatible with the concatenated alignment topology, with the mean AU probabilities ranging from 0.489 to 0.513. In a sharp contrast, the COG0012 tree topology was incompatible with the 29-marker tree, with an AU probability of 0.00023 (Supplementary Table 5, Supplementary Data File 2).

To further probe the evolutionary relationships between eukaryotes, Asgard and other archaea, a bootstrap-like subsampling was performed on the set of the 29 markers excluding COG0012. Of the 100 trees that were constructed from concatenated alignments of 14 to 23 sampled markers, 99 had the 2D topology and only one had the 3D topology (Supplementary Table 5, Supplementary Data File 2). Among these, in 62 trees, the eukaryote branch occupied different positions within the Asgard archaea, most commonly, as a sister group to the Heimdall-Gerd-Kari-Hod-Wukong clade (53 trees). In 23 trees, eukaryotes were a sister group to the Asgard-TACK clade (Supplementary Data File 2). None of the markers showed a strong association with a particular 2D tree topology akin to the dependence of the 3D topology on COG0012. The only apparent influential data point was COG0201 (Preprotein translocase subunit SecY) that was present in 40 of the 53 samples where eukaryotes grouped with the Heimdall archaea *sensu lato* and in only 6 of the 23 samples where eukaryotes were a sister branch to Asgard-TACK archaea. Not surprisingly, in the individual COG0201 tree, eukaryotes fell within the Heimdall-Gerd-Kari-Hod-Wukong clade, with a strong support (Extended Data Fig. 7c, Supplementary Data File 2). Leave-one-out analysis of all possible 29 sets of 28 genes each fully supported the 2D topology and confirmed the lack of substantial effect of any of the individual genes (Supplementary Table 5). A consensus tree of the 100 bootstrap-like marker samples and 29 leave-one-out marker samples shows

eukaryotes branching from within Asgard archaea, as a sister to the Heimdall-Gerd-Kari-Hod-Wukong clade (Fig. 1d).

The sets of archaeal, bacterial and eukaryote species employed for the tree reconstruction were selected to ensure the widest possible representation of each of the domains and therefore include, among others, groups that consist (primarily) of parasites that tend to evolve fast and could hamper accurate phylogenetic reconstruction (DPANN archaea, mycoplasma, microsporidia and several others). In phylogenetic trees, the species from these groups usually form long branches with uncertain positions. To assess the potential effect of the inclusion of these species on the domain-level tree topology, we produced the tree from a reduced species sample set that excluded highly derived parasitic clades (Supplementary Table 4). In the tree constructed from the concatenated alignment of the 29 markers (excluding COG0012), eukaryotes grouped with the Heimdall-Gerd-Kari-Hod-Wukong branch within the Asgard clade (Extended Data Fig. 7d, Supplementary Table 5, Supplementary Data File 2).

In principle, HGT from eukaryotes to Asgard archaea or from Asgard to the ancestor of eukaryotes could produce a confounding signal resulting in apparent phylogenetic affinity between eukaryotes and Asgard and biasing the concatenated trees toward the 2D topology. To assess this possibility, a tree was constructed from a concatenated alignment of the 29 markers (excluding COG0012) excluding the Asgard sequences. In this tree, the eukaryotic branch confidently grouped with the TACK superphylum, well within the diversity of the extant Archaea (Supplementary Table 5, Extended Data Fig. 7e), indicating that for this set of markers, the 2D topology is robust and is not predicated on the specific Asgard-eukaryote affinity, whether reflecting common descent or HGT.

5. Eukaryotic Signature Proteins in Asgard archaea

The computational strategy for delineating an extensive yet robust ESP set is described under Methods. The set of identified ESPs contained 505 asCOGs, including 238 that were not closely similar (E-value=10^{-10}, length coverage 75%) to those previously described by Zaremba-Niedzwiedzka et al. [2] (Supplementary Table 7). In a general agreement with previous observations, the majority of these ESPs, 329 of the 505, belonged to the 'Intracellular trafficking, secretion, and vesicular transport' (U) functional class, followed by 'Posttranslational modification, protein turnover, chaperones' (O), with 101 asCOGs (Supplementary Table 7). Among the asCOGs in the U class, 130 were Roadblock/LC7 superfamily proteins, including longins, sybindin and profilins, and 94 were small GTPases of several families, such as RagA-like, Arf-like and Rab-like ones, as discussed previously [25].

The phyletic patterns of ESP asCOGs in Asgard archaea are extremely patchy and largely lineage-specific (Extended Data Fig. 8), indicating that most of the proteins in this set are not uniformly conserved throughout Asgard evolution, but rather, are prone to frequent HGT, gene losses and duplications. Considering that the patchy distribution of the ESPs could be affected by genome incompleteness, this analysis was performed for the 76 Asgard genomes that were estimated to be at least 90% complete (see Materials and Methods for details). Capture of genes via HGT, gene loss and duplication are correlated in prokaryotes, resulting in the overall picture of dynamic evolution that is prominent in the U category COGs [26]. Even the most highly conserved ESP asCOG are missing in some Asgard lineages but show multiple duplications in others (Extended Data Fig. 8 and Supplementary Table 7).

Characteristically, many ESPs are multidomain proteins, with 37% assigned to more than one asCOG, compared to 17% among non-ESP proteins (Supplementary Table 7). Some multidomain ESPs in Asgard archaea have the same domain organizations as their homologs in eukaryotes, but these are a minority and typically contain only two domains. Examples include the fusion of two EAP30/Vps37 domains [27], and Vps23 and E2 domains [27] in ESCRT complexes, multiple Rag family GTPases, in which longin domain is fused to the GTPase domain, and several others. However, most of the domain architectures of the multidomain ESP proteins were not detected in eukaryotes and often were found only in a narrow subset of Asgard archaea, suggesting extensive domain shuffling during Asgard evolution (Fig. 3a). For example, we identified many proteins containing a fusion of Vps28/Vps23 from ESCRT I complex [27] with C-terminal domains of several homologous subunits of adaptin and COPI coatomer complexes [75,76], and E3 UFM1-protein ligase 1, which is involved in the UFM1 ubiquitin pathway [77] (Fig. 3a). Generally, a protein with such a combination of domains could be predicted to be involved in ubiquitin-dependent membrane remodeling but, because its domain architecture is unique, the precise function cannot be inferred.

The majority of the ESP genes of Asgard archaea do not belong to conserved genomic neighborhoods, but several such putative operons were detected. Perhaps, the most notable one is the ESCRT neighborhood which includes genes coding for subunits of ESCRT I, II and III, and often, components of the ubiquitin system [2], suggesting an ancient link between the two systems that persists in eukaryotes [27]. Another predicted operon is conserved in most Asgard archaea and consists of genes encoding a LAMTOR1-like protein of the Roadblock superfamily, a Rab-like small GTPase, and a protein containing the DENN (differentially expressed in normal and neoplastic cells) domain that so far has been identified only in eukaryotes (Fig. 3b). Two proteins consisting of a DENN domain fused to longin are subunits of the folliculin (FLCN) complex that is conserved in eukaryotes. The FLCN complex is the sensor of amino acid starvation interacting with Rag GTPase and Ragulator lysosomal complex, and a key component of

the mTORC1 pathway, the central regulator of cell growth in eukaryotes [78]. Some Heimdallarchaea encode several proteins with the exact same domain organization as FLCN (Fig. 3b). Ragulator is a complex that consists of 5 subunits, each containing the Roadblock domain. In Asgard archaea, however, the GTPase present in the operon is from a family that is distinct from the Rag GTPases, which interact with both FLCN and Ragulator complexes in eukaryotes, despite the fact that Rag family GTPases are abundant in Asgard archaea [25] (Supplementary Table 7). Nevertheless, this conserved module of Asgard proteins is a strong candidate to function as a guanine nucleotide exchange factor for Rab and Rag GTPases, analogously to the eukaryotic FLCN. In eukaryotes, the DENN domain is present in many proteins with different domain architectures that interact with different partners and perform a variety of functions [28,29]. The Asgard archaea also encode other DENN domain proteins, and the respective genes form expanded families of paralogs in Loki, Hel and Heimdall lineages, again, with domain architectures distinct from those in eukaryotes (Fig. 3b) [79].

Given the identification of a FLCN-like complex, a search was performed for other components of the mTORC1 regulatory pathway in Asgard archaea. The GATOR1 complex that consists of three subunits, Depdc5, Nprl2, and Nprl3, is another amino acid starvation sensor that is involved in this pathway in eukaryotes [30]. Nitrogen permease regulators 2 and 3 (NPRL2 and NPRL3) are homologous GATOR1 subunits that contain a longin domain and a small NPRL2-specific C-terminal domain [30]. We identified a protein family with this domain organization in most Thor MAGs and a few Loki MAGs. Several other ESP asCOGs include proteins with high similarity to the longin domain of NPRL2. Additionally, we identified many fusions of the NPRL2-like longin domain with various domains related to prokaryotic two-component signal transduction system (Fig. 3c). Considering the absence of a homolog of phosphatidylinositol 3-kinase, the catalytic domain of the mTOR protein, it seems likely that, in Asgard archaea, the key growth regulation pathway remains centered at typical prokaryotic two-component signal transduction systems whereas at least some of the regulators and sensors in this pathway are "eukaryotic". The abundance of NPRL2-like longin domains in Asgard archaea implies that the link between this domain and amino acid starvation regulation emerged at the onset of Asgard evolution if not earlier.

6. Reconstruction of metabolic pathways in Asgard archaea

Examination of the distribution of the asCOGs among the 12 Asgard archaeal phyla showed that the metabolic pathway repertoire was conserved among the MAGs of each phylum but differed between the phyla (Fig. 2a). Three distinct lifestyles were predicted by the asCOG analysis for different major branches of Asgard archaea, namely, anaerobic heterotrophy, facultative aerobic heterotrophy, and

chemolithotrophy (Fig. 4, Extended Data Fig. 9). For the last Asgard archaeal common ancestor (LAsCA), a mixotrophic lifestyle, including both production and consumption of H_2, can be inferred from parsimony considerations (Fig. 4, Supplementary Table 8; see Methods for further details). Loki-, Thor-, Hermod-, Baldr- and Borrarchaeota encode all enzymes for the complete (archaeal) Wood-Ljungdahl pathway (WLP) and are predicted to oxidize organic substrates, likely, by using the reverse WLP, given the lack of enzymes for oxidation of inorganic compounds (e.g., hydrogen, sulfur/sulfide and nitrogen/ammonia). The genomes of these five Asgard phyla encode homologues of membrane-bound respiratory H_2-evolving Group 4 [NiFe] hydrogenase and/or cytosolic cofactor-coupled bidirectional Group 3 [NiFe] hydrogenase [35]. Phylogenetic analysis of both group 4 and group 3 [NiFe] hydrogenases showed that Asgard archaea form distinct clades well separated from the functionally characterized hydrogenases, hampering the prediction of their specific functions in Asgard archaea (Extended Data Fig. 10a and 10b, respectively). The functionally characterized group 4 [NiFe] hydrogenases in the Thermococci are involved in the fermentation of organic substrates to H_2, acetate and carbon dioxide [80,81]. The presence of group 3 [NiFe] hydrogenases suggests that these Asgard archaea cannot use H_2 as an electron donor because they lack the enzyme complex coupling H_2 oxidation to membrane potential generation. Thus, in these organisms, bifurcate electrons from H_2 are likely to be used to support the fermentation of organic substrates exclusively [80–82].

Both Wukongarchaeota genomes (As_075 and As_085) encode a bona fide membrane-bound Group 1k [NiFe] hydrogenase that could mediate hydrogenotrophic respiration using heterodisulfide as the terminal electron acceptor [83,84] (Fig. 4, Extended Data Fig. 9 and 10c). The group 1k [NiFe] hydrogenase is exclusively found in methanogens of the order Methanosarcinales (Euryarchaeota) [66], and it is the first discovery of the group 1 [NiFe] hydrogenase in the Asgard archaea. Wukongarchaeota also encode all enzymes for a complete WLP and a putative ADP-dependent acetyl-CoA synthetase for acetate synthesis. Unlike all other Asgard archaea, Wukongarchaeota lack genes for citrate cycle and beta-oxidation. Thus, Wukongarchaeota appear to be obligate chemolithotrophic acetogens. The genomes of Wukongarchaeota were discovered only in seawater of the euphotic zone of the Yap trench (0 m and 125 m). Dissolved H_2 concentration is known to be the highest in surface seawater, where the active microbial fermentation, compared to deep sea [85], could produce sufficient amounts of hydrogen for the growth of Wukongarchaeota. Hodarchaeota, Gerdarchaeota, Kariarchaeota, and Heimdallarchaeota share a common ancestor with Wukongarchaeota (Fig. 4). However, genome analysis implies different lifestyles for these organisms. Hod-, Gerd- and Kariarchaeota encode various electron transport chain components, including heme/copper-type cytochrome/quinol oxidase, nitrate reductase, and NADH dehydrogenase, most likely, allowing the use of oxygen and nitrate as electron acceptors during aerobic and anaerobic respiration,

448 respectively [35]. In addition, Hod-, Gerd- and Heimdallarchaeota encode phosphoadenosine phosphosulfate
449 (PAPS) reductase and adenylylsulfate kinase for sulfate reduction, enabling the use of sulfate as electron
450 acceptor during anaerobic respiration. Gerd-, Heimdall-, and Hodarchaeota are only found in coastal and
451 deep-sea sedimentary environments, whereas Kariarchaeota were found also in marine water. The
452 versatile predicted metabolic capacities of these groups suggest that Hod-, Gerd- and Kariarchaeota might
453 occupy both anoxic and oxic niches. In contrast, Heimdallarchaeota appear to be able to thrive only in
454 anoxic environments.

455 **References**

456 69. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Towards functional characterization of archaeal
457 genomic dark matter. *Biochem. Soc. Trans.* **47**, 389–398 (2019).

458 70. Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the
459 prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719 (2008).

460 71. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs):
461 An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales,
462 and Methanobacteriales. *Life Basel Switz.* **5**, 818–840 (2015).

463 72. Narrowe, A. B. *et al.* Complex Evolutionary History of Translation Elongation Factor 2 and
464 Diphthamide Biosynthesis in Archaea and Parabasalids. *Genome Biol. Evol.* **10**, 2380–2393 (2018).

465 73. Crotty, S. M. *et al.* GHOST: Recovering Historical Signal from Heterotachously Evolved
466 Sequence Alignments. *Syst. Biol.* **69**, 249–264 (2020).

467 74. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**,
468 492–508 (2002).

469 75. Gomez-Navarro, N. & Miller, E. Protein sorting at the ER-Golgi interface. *J. Cell Biol.* **215**, 769–
470 778 (2016).

471 76. Kibria, K. M. K. *et al.* A genome-wide analysis of coatomer protein (COP) subunits of
472 apicomplexan parasites and their evolutionary relationships. *BMC Genomics* **20**, 98 (2019).

473 77. Wei, Y. & Xu, X. UFMylation: A Unique & Fashionable Modification for Life. *Genomics*
474 *Proteomics Bioinformatics* **14**, 140–146 (2016).

475 78. Lawrence, R. E. *et al.* Structural mechanism of a Rag GTPase activation checkpoint by the
476 lysosomal folliculin complex. *Science* **366**, 971–977 (2019).

477 79. Marat, A. L., Dokainish, H. & McPherson, P. S. DENN domain proteins: regulators of Rab
478 GTPases. *J. Biol. Chem.* **286**, 13791–13800 (2011).

479 80. Yu, H. *et al.* Structure of an Ancient Respiratory System. *Cell* **173**, 1636-1649.e16 (2018).

81. Schut, G. J., Boyd, E. S., Peters, J. W. & Adams, M. W. W. The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. *FEMS Microbiol. Rev.* **37**, 182–203 (2013).

82. Bryant, F. O. & Adams, M. W. Characterization of hydrogenase from the hyperthermophilic archaebacterium, Pyrococcus furiosus. *J. Biol. Chem.* **264**, 5070–5079 (1989).

83. Deppenmeier, U., Blaut, M., Schmidt, B. & Gottschalk, G. Purification and properties of a F420-nonreactive, membrane-bound hydrogenase from Methanosarcina strain Gö1. *Arch. Microbiol.* **157**, 505–511 (1992).

84. Thauer, R. K. *et al.* Hydrogenases from methanogenic archaea, nickel, a novel cofactor, and H2 storage. *Annu. Rev. Biochem.* **79**, 507–536 (2010).

85. Conrad, R. & Seiler, W. Methane and hydrogen in seawater (Atlantic Ocean). *Deep Sea Res. Part Oceanogr. Res. Pap.* **35**, 1903–1917 (1988).