nature portfolio

https://doi.org/10.1038/s41586-025-08955-7

Supplementary information

Deep origin of eukaryotes outside Heimdallarchaeia within Asgardarchaeota

In the format provided by the authors and unedited

Deep origin of eukaryotes outside Heimdallarchaeia within Asgardarchaeota

Jiawei Zhang[1], Xiaoyuan Feng[2], Meng Li[2], Yang Liu[2], Min Liu[3], Li-Jun Hou[1]*, Hong-Po Dong[1]*

[1] State Key Laboratory of Estuarine and Coastal Research, Yangtze Delta Estuarine Wetland Ecosystem Observation and Research Station, Ministry of Education, East China Normal University, Shanghai 200241, China; [2]Archaeal Biology Center, Synthetic Biology Research Center, Shenzhen Key Laboratory of Marine Microbiome Engineering, Key Laboratory of Marine Microbiome Engineering of Guangdong Higher Education Institutes, Institute for Advanced Study, Shenzhen University, Shenzhen 518060, China; [3]Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai 200241, China

* Corresponding author: hpdong@sklec.ecnu.edu.cn (Hong-Po Dong), ljhou@sklec.ecnu.edu.cn (Li-Jun Hou).

Supplementary Information

Supplementary Tables are available as separate excel files

Supplementary Methods

1. Identification and analysis of eukaryotic signature proteins (ESPs)

The ESPs proteins were identified by using HMMER[1], diamond-BLAST[2] (e-value <1e−5), and InterproScan[3]. In the HMMER[1] identification, a bit score gathering (GA) threshold was used for each Pfam[4] family. All identified proteins were verified by searching against nr database using BLASTP[2]. The identification of these ESPs that is previously reported to be specific to Hodarchaeales is detailed as follows: 1) ε DNA polymerase Pol2. The eukaryotic protein contains three domains: PF03104, PF00136, and PF10108. To identify its homologs in eukaryotes, the coexistence of all three domains was required. The resulting sequences were manually verified by searching against the SwissProt database[5]. For Asgard archaea, homologs with any one of the three domains were considered. All identified sequences were used to infer a phylogeny using FastTree2[6] under the LG + CAT model. These Asgard archaeal sequences that harbored the typical two or three domains of the Pol2 subunit and were closely related to eukaryotic sequences are considered eukaryotic orthologs. 2) Canonical eukaryotic-type elongation factor 2 (EF-2). Homologous sequences of EF-2 were retrieved by comparison with COG0480 sequences of EF-2 using BLASTP. The resulting sequences were aligned with reference sequences of archaeal aEF-2 and eukaryotic eEF-2 from a previous study[7] using mafft-linsi[8]. Two conserved motifs in canonical EF-2 (HxDxxHRG and SPHKHN) were checked. 3) Proteins for diphthamide biosynthesis (Dph1, Dph5, and Dph6). Homologs of Dph1, Dph5, and Dph6 were identified using their respective domains: PF01866, IPR004551, and PF01902. A search for Swiss-Prot database[5] was applied to confirm these homologs. 4) N-terminal histone tails. Sequences homologous to eukaryotic histones were retrieved by searching against histone reference sequences from a previous study[9] using BLASTP[2]. The resulting sequences were verified by comparison with nr database and then aligned with reference sequences using MEGA X[10]. The N-terminal histone tails were manually checked in the alignments.

Supplementary results and discussion

1. Phylogenetic placement of Njordarchaeales and its impact on evolutionary history of eukaryotes

Although the site-exclusion treatments conducted on the S67 dataset indicated that the clustering of Njordarchaeales and Korarchaeota was not an artefact stemming from rate heterogeneity of the dataset, similar treatment applied to the NM57 dataset also strongly support the monophyly of Njordarchaeales and Heimdallarchaeia. In this case, we doubted that these Njordarchaeales MAGs were misbinned and may contain high levels of contamination which cannot be detected by CheckM[11]. CheckM[11] can generate a false sense of bin accuracy, as evidenced by integrating two partial single-cell genome bins[12, 13]. We determined the taxonomic profiles of assembled sequences of Njordarchaeales MAGs by searching against TACK and Asgard archaeal sequences from GTDB (r220) database[14] excluding Njordarchaeales-related sequences. Between 42% and 49% of the genes were classified as TACK archaea, while 19% to 37% were assigned to Asgard archaea. Subsequently, we searched the source of these Njordarchaeales MAGs and found multiple metagenomes at the sampling locations[15-17]. Reads from these metagenomes were mapped to the ten Njordarchaeales representative MAGs. Among these, contigs of four MAGs had positive hits. However, contigs within each MAG exhibited differential coverage patterns across different metagenomic samples (Fig. 2c and d) and were split into two or four clusters. The findings suggest that these MAGs may encompass genomic fragments derived from two to four distinct microbial populations. The NM57 markers were selected based on their conservation in Asgard archaea[15], thus they may harbor stronger phylogenetic signals that drag Njordarchaeales into Asgard phylum, compared to markers of other datasets. Relative to the NM57 marker proteins, ribosomal proteins may be more conserved and have fewer events of gene duplication, transfer and loss[18]. Among the 12 marker datasets analyzed, all except for the NM57 and NM200 contained varying quantities of ribosomal proteins (Supplementary Table 4). Similar to the NM200 tree, phylogenetic analyses of these

datasets containing ribosomal proteins support grouping of Njordarchaeales and Korarchaeota within the TACK superphylum (Supplementary Fig. 2). Meanwhile, in the case of individual MAG, 69-84% of ribosomal proteins were classified into the TACK superphylum (Supplementary Fig. 4). The results suggest that main bodies of Njordarchaeales genomes may be affiliated with TACK superphylum. When using different marker sets to infer phylogeny, the shift of Njordarchaeales between the TACK superphylum and Asgard phylum may be attributed to the distribution of these markers across the two parts of Njordarchaeales genomes and the intensity of their phylogenetic signals.

Eme et al[15]. used the NM57 marker set to build a supermatrix for 175 Asgard archaea and 14 eukaryotic taxa. ML tree of the untreated NM57 supermatrix supported grouping of eukaryotes with Njordarchaeales. When site-by site desaturation method was applied to the supermatrix, it is observed that the support for the monophyly of eukaryotes and Njordarchaeales dramatically dropped while support for the grouping of eukaryotes with Hodarchaeales never emerged, as the fastest-evolving sites were progressively removed. The results indicate that the presence of Njordarchaeales strongly influences the position of eukaryotes in Asgard archaea. In the dataset, partial Asgard archaeal sequences in Njordarchaeales genomes may be artefactually attracted into Asgard phylum by eukaryotes.

2. Phylogenetic position of eukaryotes in Asgard archaea

Using markers with horizontal gene transfer events may result in phylogenetic reconstruction artefacts. To enhance the robustness of our phylogenies, we added 93 representative bacterial genomes (Supplementary Table 3) to our archaeal dataset. For the S67, S97, and S150 marker sets, individual-gene trees were inferred to manually inspect any instances of inter-domain horizontal gene transfers, which can be accessed on Figshare (https://figshare.com/s/6e523322b0b647b91dda). These markers that failed to recover archaea and bacteria as reciprocally monophyletic domains were removed. By doing so, we obtained 47, 60, and 99 markers (S47, S60, and S99) from the S67, S97 and S150 marker sets, respectively (Supplementary Table 4). Based on

these newly selected markers, we constructed three supermatrices: exS47, exS60, and exS99, using 411 Asgard archaea and 14 eukaryotes. ML phylogenomic analyses of the three supermatrices under the PMSF approximation of LG + C60 + F + G placed eukaryotes as a sister group to Heimdallarchaeia (Supplementary Fig. 9), which is in accordance with the tree topologies inferred using the S67 and S97 datasets. We noticed that the tree topology inferred from the exS99 dataset closely resembled those inferred from site-excluded subsets of the S150 dataset, suggesting that for large datasets, either removing markers associated with inter-domain horizontal gene transfer events or excluding fast-evolving sites from alignments can effectively mitigate the impact of non-phylogenetic signals, such as rate or compositional biases, on phylogenetic inference.

For comparison, the three sets of markers (S47, S60, and S99) were also used to infer ML phylogenies from the genomic data used by Eme et al.[15] (175 Asgard archaea + 14 eukaryotes), under the PMSF approximation of LG + C60 + F + G. The resulting supermatrices were named smS47, smS60, and smS99, respectively. These analyses supported the Heimdallarchaeia-sister topology (Supplementary Fig. 10). Furthermore, we investigated the influence of rate heterogeneity on the placement of eukaryotes by progressively reducing fastest-evolving sites of these alignments and inferring ML phylogenies. For the smS47 dataset, the support for Heimdallarchaeia-sister was unstable during the removal of fast-evolving sites, with significant support emerging only after excluding 20% of the fastest-evolving sites. No support for Hodarchaeales-sister was observed (Supplementary Fig. 11a). For the smS60 and smS99 datasets, the support for Heimdallarchaeia-sister remained robust until 50% and 80% of sites were excluded, respectively (Supplementary Fig. 11b and c). The results suggest that larger datasets retain more phylogenetic information for resolving the placement of eukaryotes.

Similarly, the site-exclusion method was applied to the exS47, exS60, and exS99 supermatrices (Supplementary Fig. 12). We observed that in the exS47 and exS60 supermatrices, the loss of phylogenetic signals supporting the Heimdallarchaeia-sister

relationship occurred at a significantly slower rate following the removal of the fastest-evolving sites, compared to the corresponding supermatrices constructed using Eme et al.'s data. More importantly, as showed in Fig. 3c, when the NM57 marker set selected by Eme et al. was used to construct a supermatrix for our expanded Asgard archaeal sampling, phylogenomic analyses of subsets of this supermatrix strongly supported the Heimdallarchaeia-sister topology after the exclusion of fast-evolving sites. These results indicate that our expanded Asgard archaeal sampling helps to disentangle true phylogenetic signals from noise and potential artefacts.

The different patterns observed during the removal of fast-evolving sites between the NM57 dataset, which lacks ribosomal protein markers, and the datasets containing ribosomal protein markers likely suggest that datasets generated using functional proteins as markers may contain a higher level of rate heterogeneity across sites. It remains unclear whether this is related to ribosomal proteins being more evolutionarily conserved than functional proteins[18]. However, identifying more vertically inherited markers to mitigate the challenges in phylogenetic tree construction caused by long branches, such as those of eukaryotes, remains an important task for the future.

3. Distribution of eukaryotic signature proteins (ESPs) in the expanded genomic diversity of Asard archaea

Eme et al[15]. reported that some ESPs such as the ε DNA polymerase catalytic subunit (Pol2), eukaryotic-type EF2, proteins for the diphthamide biosynthesis, RNA polymerase subunit RPB8 and amino-terminal histone tails are specific to Hodarchaeales or Njordarchaeales, and argued that the presence of these ESPs in Hodarchaeales may support close relationship between Hodarchaeales and eukaryotes. Nevertheless, we examined the distribution of these ESPs in the expanded sampling of Asgard archaea and found that they were present in multiple clades of Asgard archaea including Asgard lineages outside or inside Heimdallarchaeia (Supplementary Fig. 13, Supplementary Table 9). For example, we identified homologs of eukaryotic ε DNA polymerase Pol2 in 13 Asgard clades by using

HMMER[19] with a GA bit score cutoff and inferring a phylogeny. Phylogenetic analysis revealed that the Pol2 sequences of eukaryotes formed a distinct cluster with those of Jordarchaeia, Wenzhongarchaeales, and Hodarchaeales (designated as Cluster 1). In addition, two other clusters (Cluster 2 and 3), closely related to Cluster 1, contain sequences from multiple Asgard lineages (excluding Hodarchaeales), all of which harbored the typical two or three domains of the Pol2 subunit (Supplementary Fig. 14).

Eme et al. argued that the simultaneous presence of *dhp* genes for the diphthamide biosynthesis and canonical eukaryotic EF-2 gene is unique to Hodarchaeales MAGs. Using BLASTP[2] and HMMER[19] in combination with sequence alignment, we also identified genes encoding canonical EF-2[7] in addition to *dhp* genes in members of Baldrarchaeia, Jordarchaeia, and Asgardarchaeia (Supplementary Fig. 15). A comparison with archaeal and eukaryotic histone sequences from a previous study[9], using BLASTP, revealed that N-terminal histone tails are also present in histones from members of Heimdallarchaeaceae, Helarchaeales, Lokiarchaeales, and Thorarchaeia (Supplementary Fig. 16). Finally, for the RNA polymerase subunit RPB8, using the GA bit score cutoff for the PF16992/PF03870 domain of the protein, no hits were retrieved from members of Hodarchaeales or other lineages of Heimdallarchaeia. However, it could be identified in members of Baldrarchaeia, Helarchaeales, Jordarchaeia, Odinarchaeia, and Sifarchaeia (Supplementary Fig. 13).

4. Implications for the discovery of the Njordarchaeales chimerism

Reconstructing genomes from environmental metagenome is a key technique for uncovering the mysteries of unculturable microorganisms on earth. However, the chimeric nature of the Njordarchaeales genomes that we discovered serves as a cautionary reminder for the genome reconstruction, highlighting that these assembled genomes may not always be entirely accurate. Without thorough inspection, newly generated genomes can lead to misinterpretations of microbial evolution and ecological functions. To mitigate this, we recommend that for newly discovered

microbial lineages, multiple reliable marker datasets should be employed to analyze their evolutionary relationships with other microorganisms. If the evolutionary positions of these lineages appear unstable in the tree of life, state-of-the-art classification tools should be used to determine the taxonomic profiles of their contigs. Furthermore, sequences of these new lineages should be retrieved from different metagenomic samples to analyze their sequence features. Genomes that exhibit distinct sequence characteristics across different metagenomes should be approached with caution.

Supplementary references

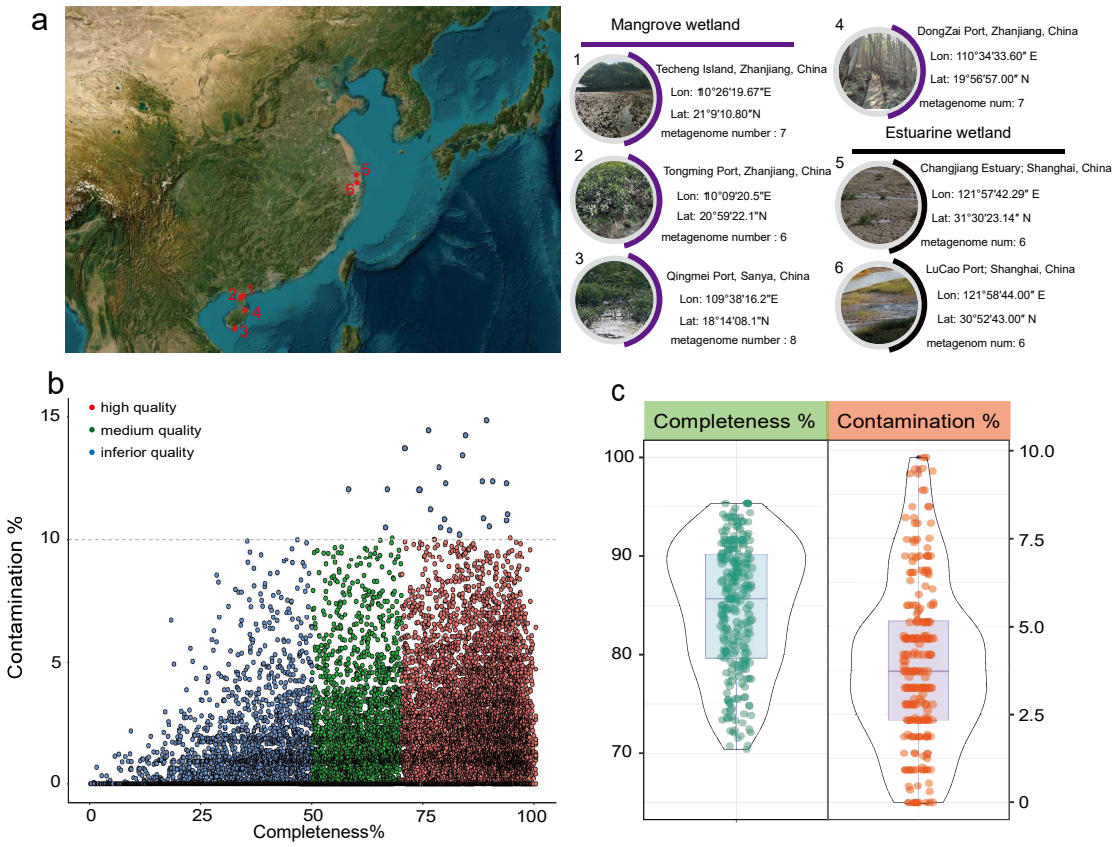
1. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
2. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59-60 (2015).
3. Jones, P. et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
4. Finn, R.D. et al. Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222-D230 (2014).
5. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45-48 (2000).
6. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
7. Narrowe, A.B. et al. Complex Evolutionary History of Translation Elongation Factor 2 and Diphthamide Biosynthesis in Archaea and Parabasalids. *Genome Biol. Evol.* **10**, 2380-2393 (2018).
8. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).
9. Henneman, B., van Emmerik, C., van Ingen, H. & Dame, R.T. Structure and function of archaeal histones. *PLoS Genet.* **14**, e1007582 (2018).
10. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547-1549 (2018).
11. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P. & Tyson, G.W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043-1055 (2015).

12. Becraft, E.D. et al. Rokubacteria: Genomic Giants among the Uncultured Bacterial Phyla. *Front. Microbiol.* **8**, 2264 (2017).
13. Chen, L.X., Anantharaman, K., Shaiber, A., Eren, A.M. & Banfield, J.F. Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315-333 (2020).
14. Parks, D.H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785-D794 (2022).
15. Eme, L. et al. Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* (2023).
16. Dombrowski, N., Seitz, K.W., Teske, A.P. & Baker, B.J. Genomic insights into potential interdependencies in microbial hydrocarbon and nutrient cycling in hydrothermal sediments. *Microbiome* **5**, 106 (2017).
17. Zhou, Z., St John, E., Anantharaman, K. & Reysenbach, A.L. Global patterns of diversity and metabolism of microbial communities in deep-sea hydrothermal vent deposits. *Microbiome* **10**, 241 (2022).
18. Ramulu, H.G. et al. Ribosomal proteins: toward a next generation standard for prokaryotic systematics? *Mol. Phylogenet Evol.* **75**, 103-117 (2014).
19. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
20. Spang, A. et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173-179 (2015).
21. Zaremba-Niedzwiedzka, K. et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353-358 (2017).
22. Da Cunha, V., Gaia, M., Gadelle, D., Nasir, A. & Forterre, P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).
23. Dombrowski, N. et al. Undinarchaeota illuminate DPANN phylogeny and the impact of gene transfer on archaeal evolution. *Nat. Commun.* **11**, 3939 (2020).
24. Williams, T.A., Cox, C.J., Foster, P.G., Szollosi, G.J. & Embley, T.M.

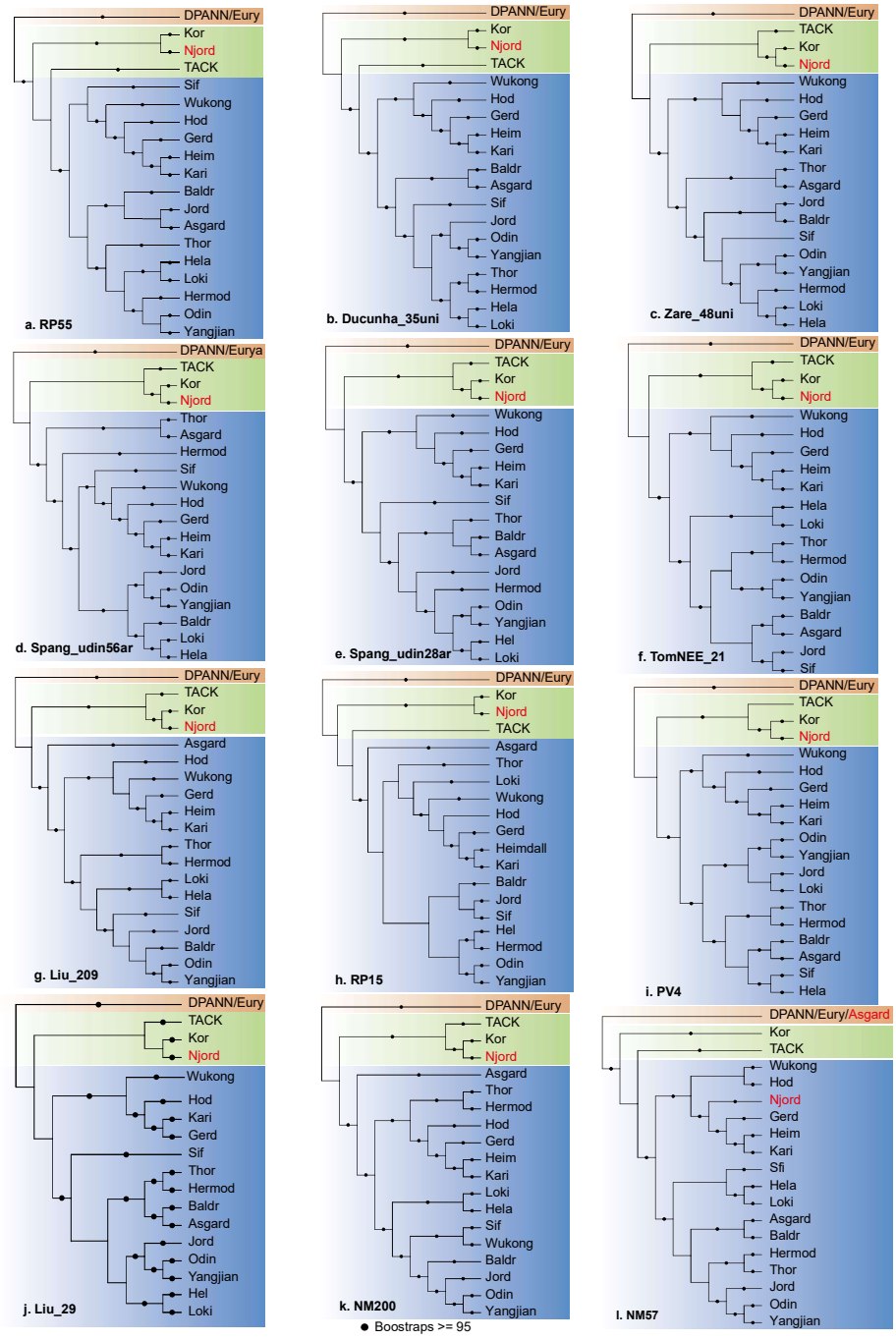
Phylogenomics provides robust support for a two-domains tree of life. *Nat. Ecol. Evol.* **4**, 138-147 (2020).

25. Liu, Y. et al. Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553-557 (2021).
26. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T.J.G. Deep mitochondrial origin outside the sampled alphaproteobacteria. *Nature* **557**, 101-105 (2018).
27. Darling, A.E. et al. PhyloSift: phylogenetic analysis of genomes and metagenomes. *Peerj* **2**, e243 (2014).
28. Petitjean, C., Deschamps, P., Lopez-Garcia, P., Moreira, D. & Brochier-Armanet, C. Extending the conserved phylogenetic core of archaea disentangles the evolution of the third domain of life. *Mol. Biol. Evol.* **32**, 1242-1254 (2015).
29. Lyons, T.W., Reinhard, C.T. & Planavsky, N.J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307-315 (2014).

Supplementary Figures

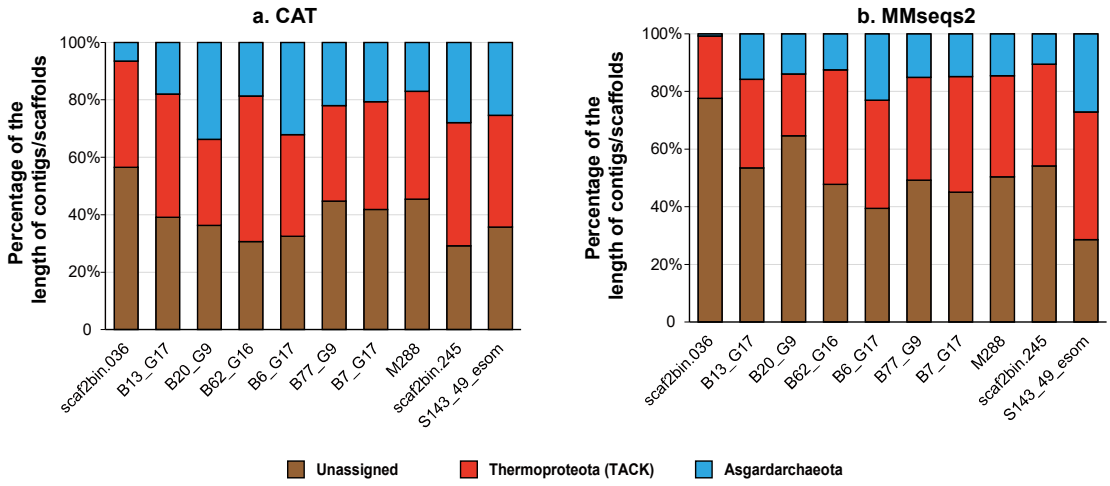


Supplementary Figure 1. Sampling locations and quality of metagenome-assembled genomes. (a) The map illustrates the sampling locations of sediment cores and detailed information of sampling was provided in the plot on the right; (b) The scatter plot showing the quality of 11,878 genomes recovered from 40 sediment samples. The completeness and contamination of these genomes were determined by CheckM; (c) Quality of 223 Asgard genomes generated in this study. The median completeness and contamination values are 85.67% and 3.74%, respectively, indicating that these genomes are nearly complete. The map presented in panel a was generated using the leaflet package in R.

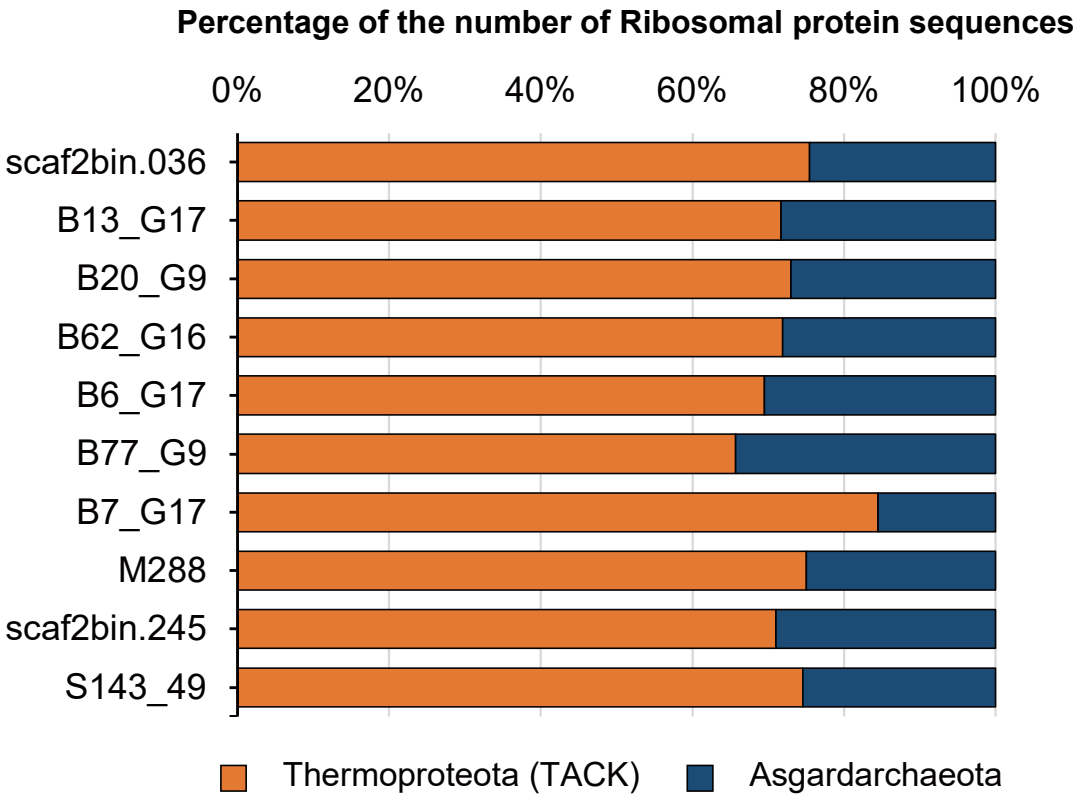


Supplementary Figure 2. Phylogenomic analysis of 12 frequently used archaeal marker sets. The maximum likelihood trees were constructed using IQ-TREE with the LG+C60+F+G+PMSF model, based on a dataset comprising 579 archaeal taxa (411 Asgard archaea, 51 DPANN archaea, 47 Euryarchaea, and 70 representatives of TACK archaea). Bootstrap support values ≥ 95% are represented by black dots. The marker sets used include RP55[20, 21] (55 proteins, 6,295 sites), DaCunha_35uni[22] (35

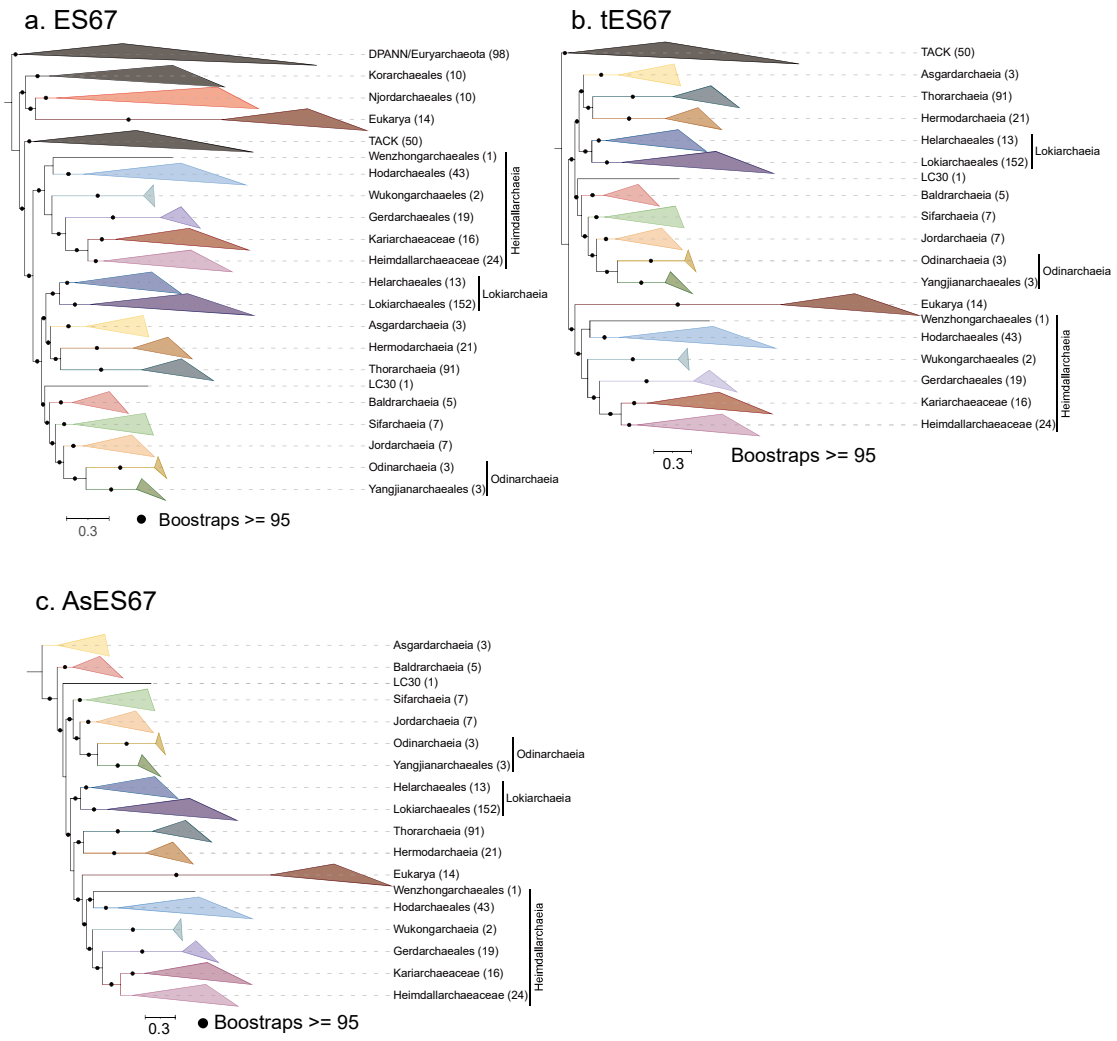
proteins, 7,790 sites), Zare_48uni[21] (48 proteins, 9,840 sites), Spang_udin56ar[23] (56 proteins, 12,608 sites), Spang_udin28ar[23] (28 proteins, 7,579 sites), TomNEE_21uni[24] (21 proteins, 6,084 sites), Liu_209[25] (209 proteins, 35,422 sites), RP15[26] (15 proteins, 2,174 sites), PV4[27] (37 proteins, 6,346 sites), NM200[28] (200 proteins, 47,475 sites) and NM57[15] (57 proteins, 13,973 sites). Eury, Euryarchaeota; Kor, Korarchaeia; Njord, Njordarchaeales; Wukong, Wukongarchaeia; Hod, Hodarchaeales; Gerd, Gerdarchaeales; Heim, Heimdallarchaeaceae; Kari, Kariarchaeaceae; Baldr, Baldrarchaeia; Asgard, Asgardarchaeia; Sif, Sifarchaeia; Jord, Jordarchaeia; Odin, Odinarchaeia; Yangjian, Yangjianarchaeales; Thor, Thorarcheaia; Hermod, Hermodarhcaeia; Hela, Helarchaeales; Loki, Lokiarchaeales.



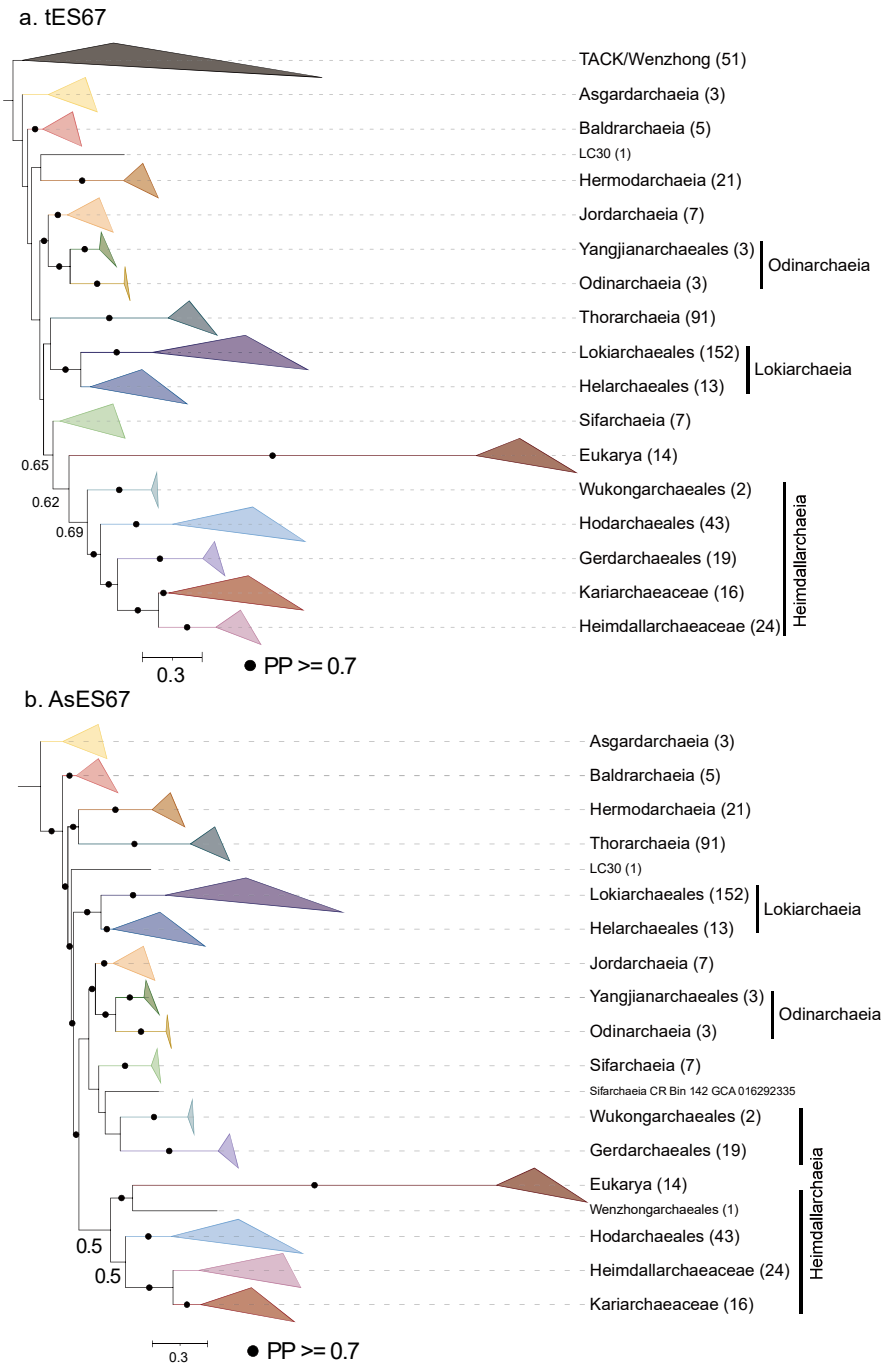
Supplementary Figure 3. Taxonomic profiles of contigs/scaffolds in the Njordarchaeales genomes. The percentage of the length of contigs/scaffolds assigned to Thermoproteota or Asgardarchaeota based on sequence length for each of the 10 Njordarchaeales representatives was determined using the (a) CAT and (b) MMseq2.



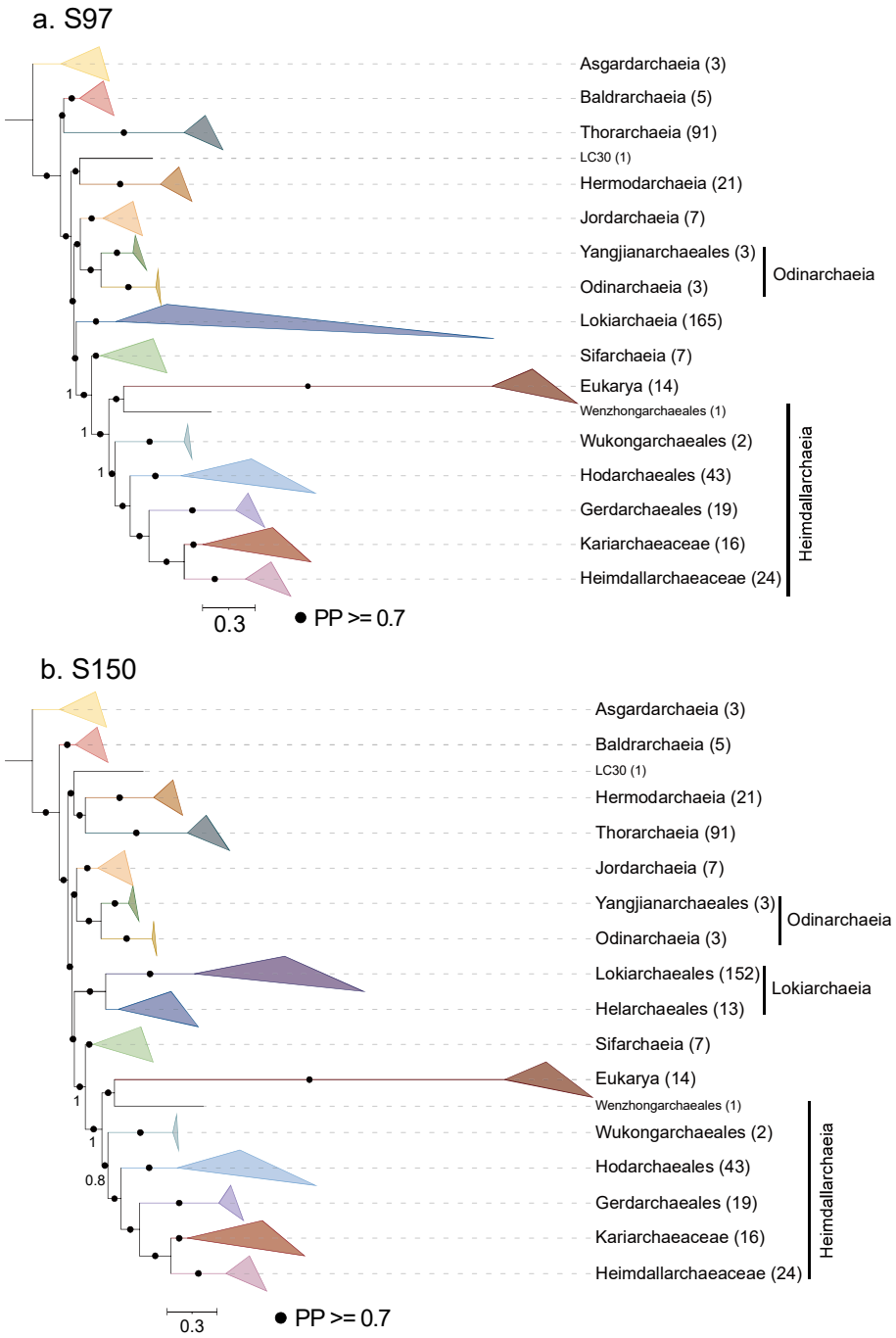
Supplementary Figure 4. Percentage of ribosomal proteins assigned to Thermoproteota (TACK superphylum) or Asgardarchaeota, based on the count for each of the 10 Njordarchaeales representatives.



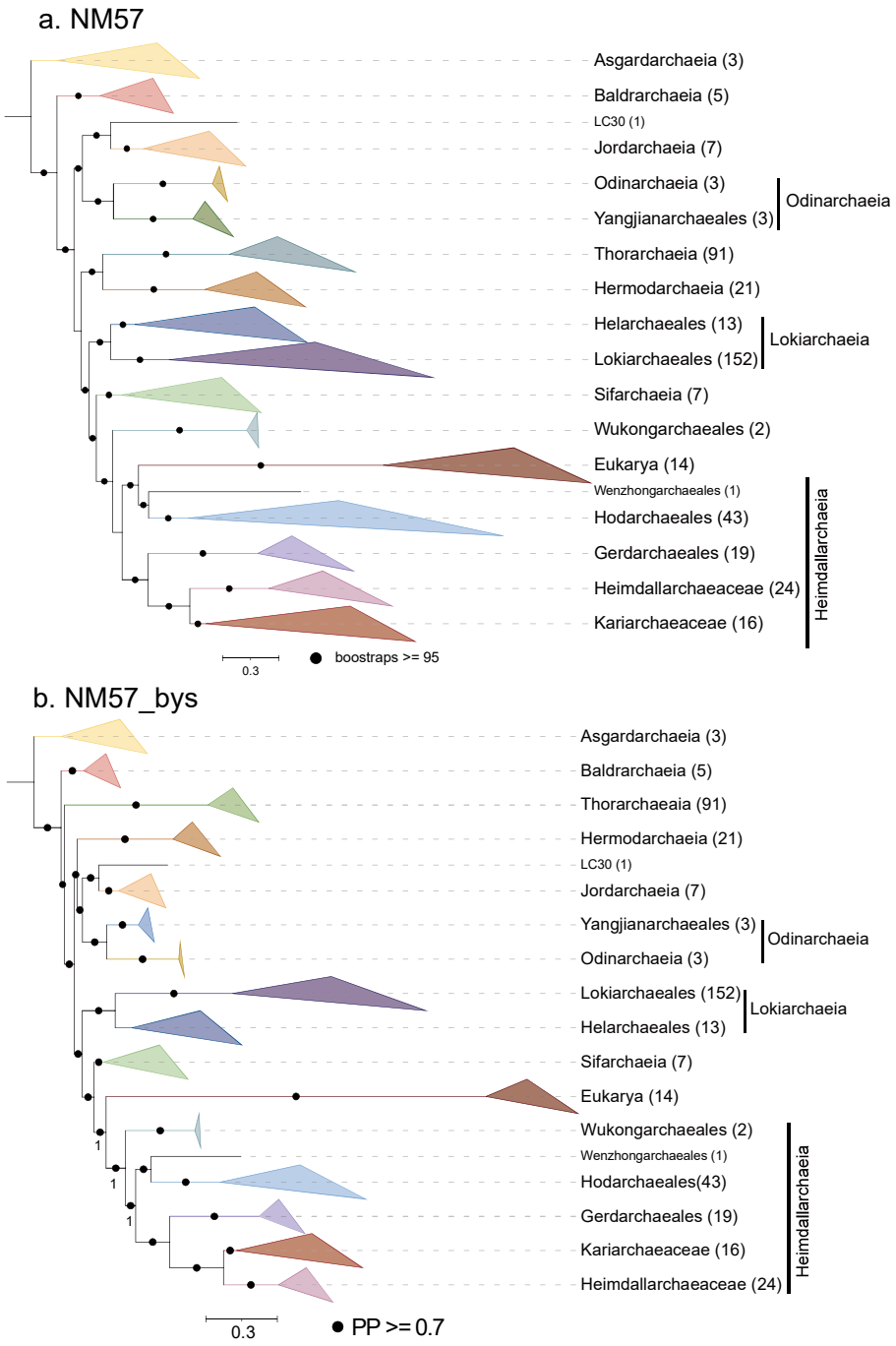
Supplementary Figure 5. The maximum likelihood phylogenomic analysis of multiple datasets, showing the placement of eukaryotes. (a) ES67 dataset (67 concatenated proteins, 593 taxa, 11,596 sites). (b) tES67 dataset (67 concatenated proteins, 475 taxa, 12,720 sites). (c) AsES67 dataset (67 concatenated proteins, 425 taxa, 13,348 sites). The trees were inferred using IQ-TREE under the LG+C60+F+G+PMSF model. Bootstrap support values ≥ 95% are represented by black dots. The scale bar denotes the average expected number of substitutions per site.



Supplementary Figure 6. Bayesian phylogenetic tree of the tES67 and AsES67 datasets, showing the position of eukaryotes. (a) tES67 dataset (12,720 sites, 475 taxa, 2 chains, 50000 generations). (b) AsES67 dataset (13,348 sites, 425 taxa, 2 chains, 35000 generations). Bayesian inferences were performed using PhyloBayes (v1.9) MPI with the CAT +GTR after the alignments were recoded into four categories using the SR4 scheme. PP, posterior probability support.

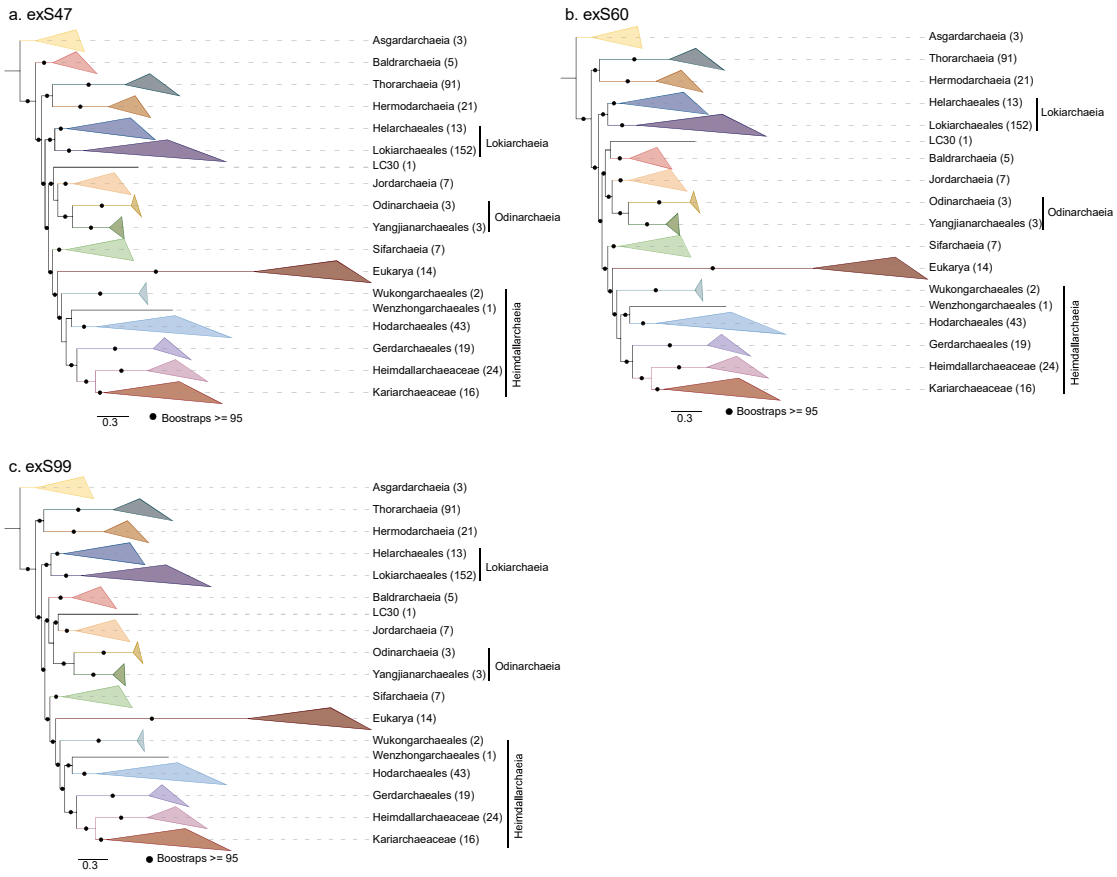


Supplementary Figure 7. Bayesian phylogenetic tree of the S97 and S150 datasets. (a) S97 dataset (12,720 sites, 425 taxa, 2 chains, 20000 generations). (b) S150 dataset (32,277 sites, 425 taxa, 2 chains, 20000 generations). Bayesian inferences were performed using PhyloBayes (v1.9) MPI with the CAT + GTR, based on 411 Asgard archaeal taxa and 14 eukaryotic taxa. Prior to analysis, the alignments were recoded into four categories using the SR4 scheme.

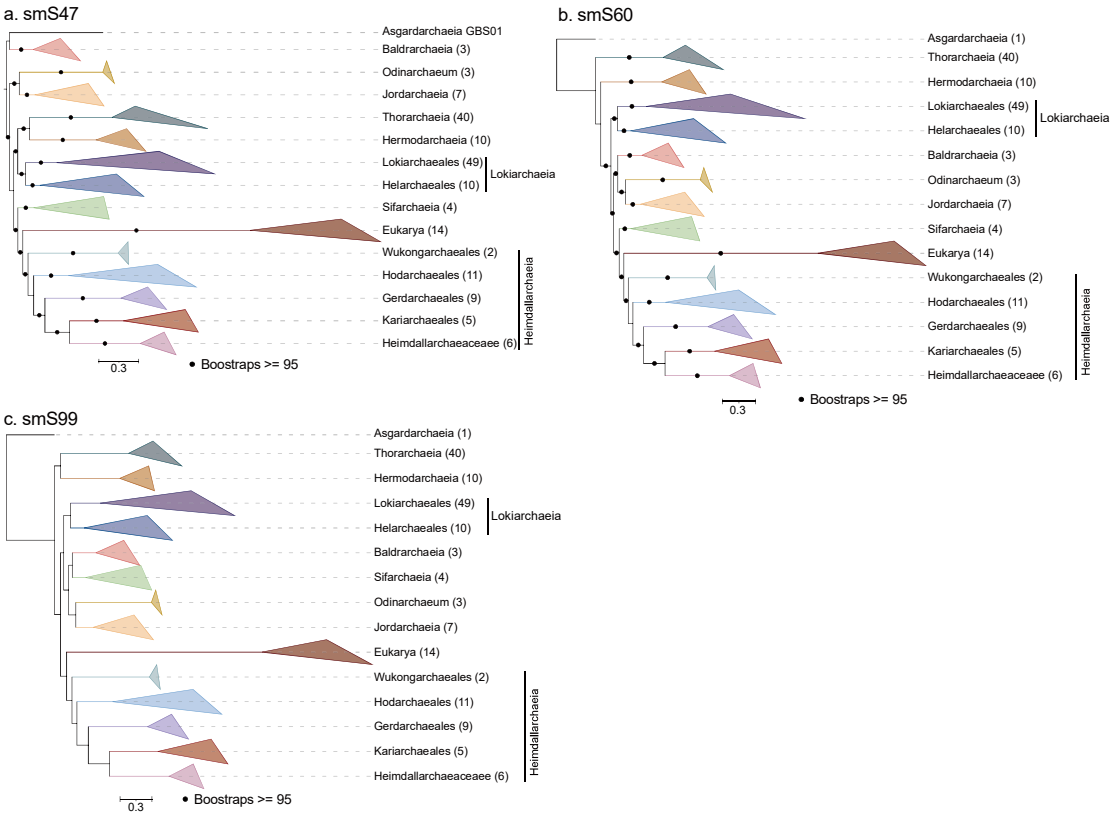


Supplementary Figure 8. Phylogenetic analyses of the NM57 dataset based on 411 Asgard archaeal taxa and 14 eukaryotic taxa. (a) The maximum likelihood phylogenomic analysis based on the NM57 dataset (15,444 sites, 57 concatenated proteins, 425 taxa). The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model. Bootstrap support values ≥ 95% are represented by black dots. The scale bar denotes the average expected number of substitutions per site; (b) Bayesian

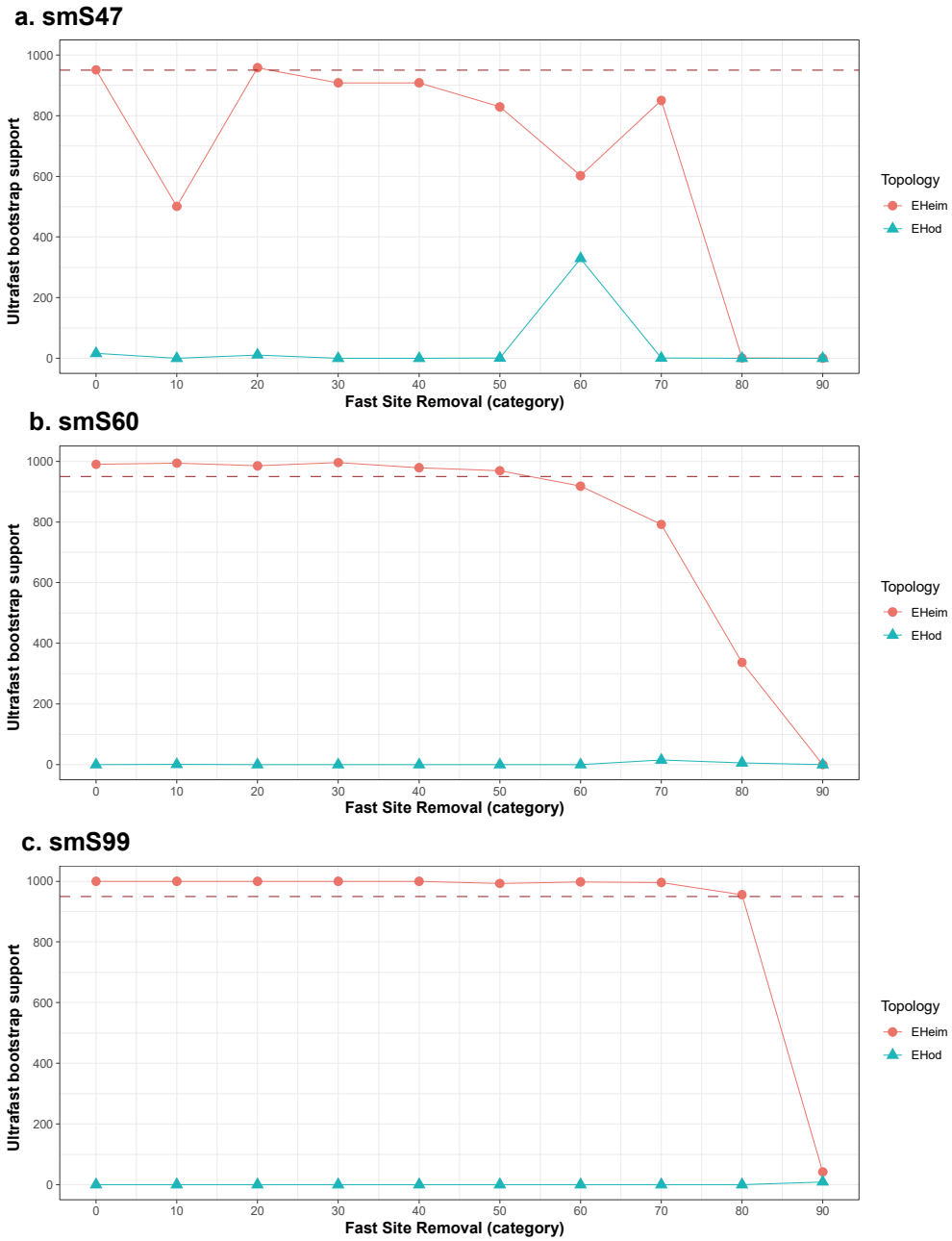
phylogenetic tree of the NM57 dataset (15,444 sites, 425 taxa, 2 chains, 20000 generations). Bayesian inferences were performed using PhyloBayes (v1.9) MPI with the CAT + GTR after the alignments were recoded into four categories using the SR4 scheme. PP, posterior probability support.



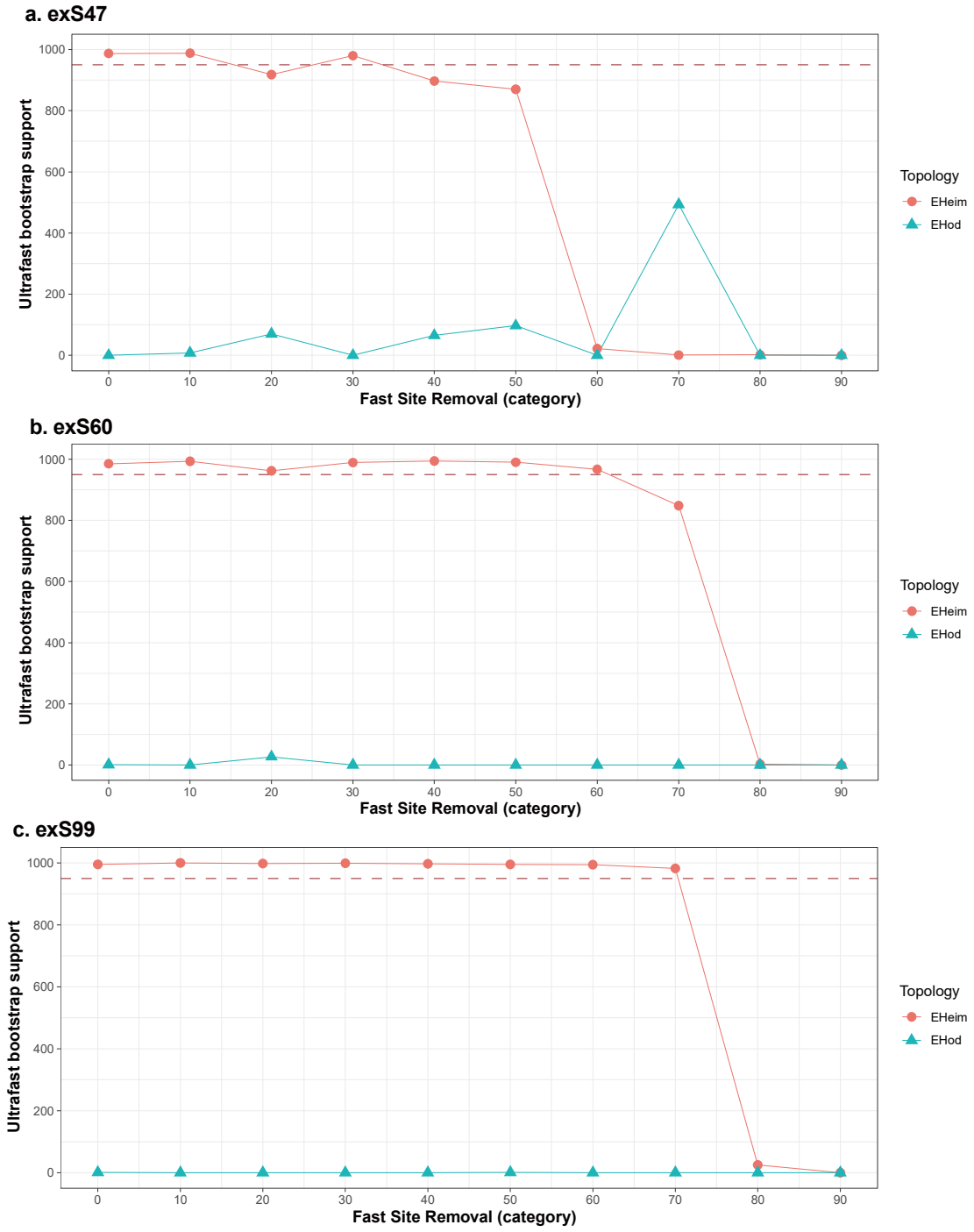
Supplementary Figure 9. The maximum likelihood phylogenomic analysis of the exS47, exS60 and exS99 datasets. (a) exS47 dataset (47 concatenated proteins, 425 taxa, 9291 sties), (b) exS60 dataset (60 concatenated proteins, 425 taxa, 11932 sties), (c) exS99 dataset (99 concatenated proteins, 425 taxa, 20547 sties). The trees were inferred using IQ-TREE under the LG+C60+F+G+PMSF model. Bootstrap support values ≥ 95% are represented by black dots. The scale bar denotes the average expected number of substitutions per site.



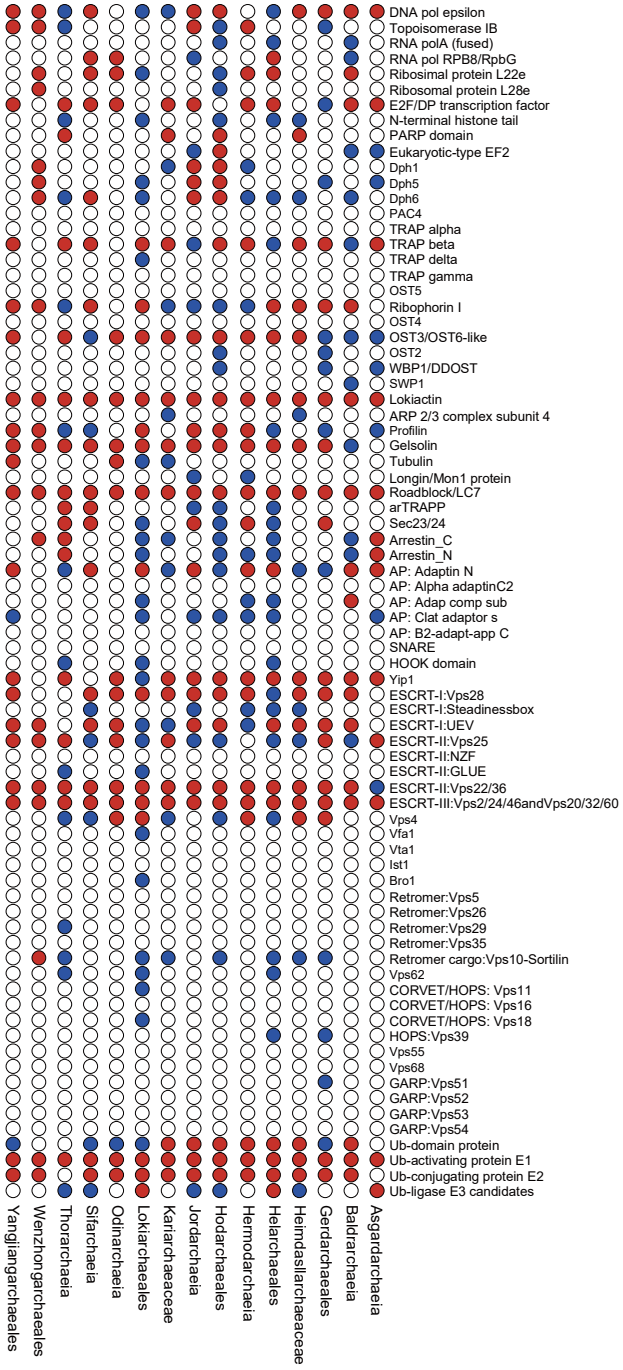
Supplementary Figure 10. The maximum likelihood phylogenomic analysis of the smS47, smS60 and smS99 datasets. (a) smS47 (47 concatenated proteins, 174 taxa, 9300 sties), (b) smS60 dataset (60 concatenated proteins, 174 taxa, 11910 sties), (c) smS99 dataset (99 concatenated proteins, 174 taxa, 20594 sties). The trees were inferred using IQ-TREE under the LG+C60+F+G+PMSF model. Bootstrap support values ≥ 95% are represented by black dots. The scale bar denotes the average expected number of substitutions per site.



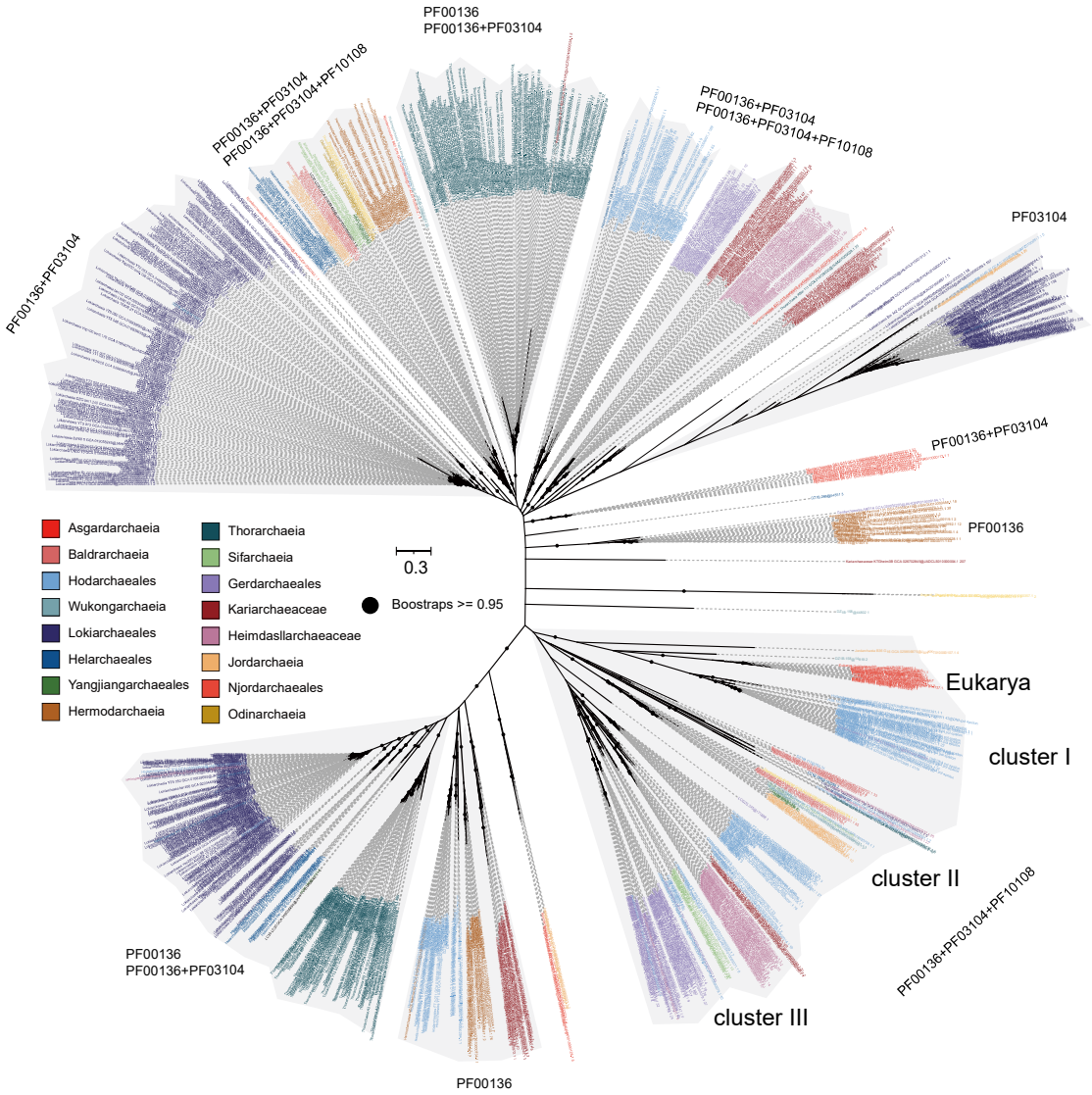
Supplementary Figure 11. Evolution of ultrafast bootstrap support for the monophyly of either eukaryotes and Heimdallarchaeia (EHeim, red line) or the monophyly of eukaryotes and Hodarchaeales (EHod, blue line). The bootstrap values were obtained from the phylogenies inferred from the smS47 dataset (a), smS60 dataset (b), and smS99 dataset (c), as the fastest-evolving sites were progressively removed. The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model.



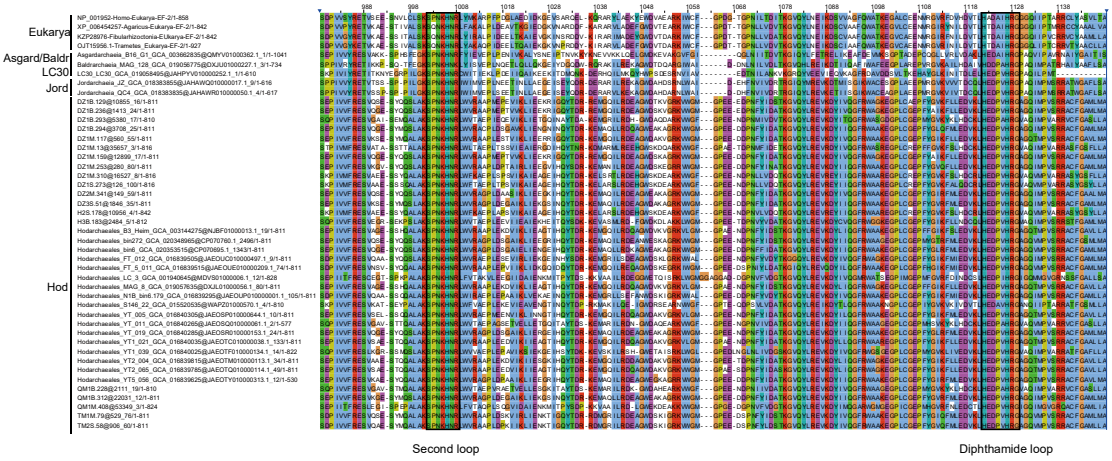
Supplementary Figure 12. Site-exclusion treatments of the exS47 (a), exS60 (b), and exS99 (c) datasets. The evolution of ultrafast bootstrap support for the monophyly of either eukaryotes and Heimdallarchaeia (EHeim, red line) or the monophyly of eukaryotes and Hodarchaeales (EHod, blue line) was tracked as the fastest-evolving sites were progressively removed. The trees were inferred using IQ-TREE under the LG + C60 + F + G + PMSF model.



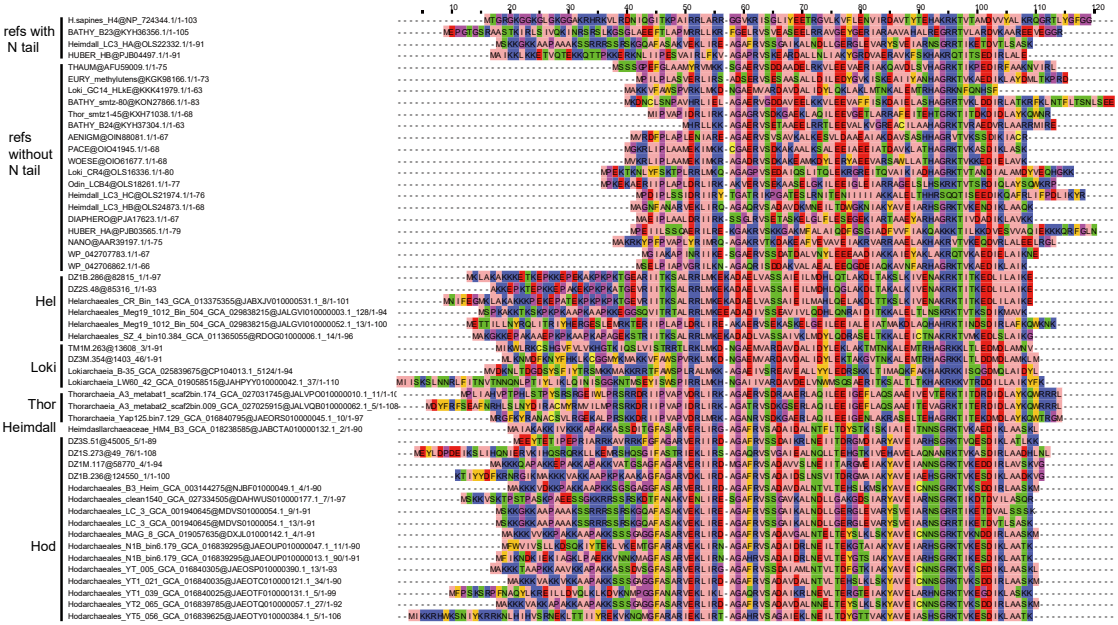
Supplementary Figure 13. Eukaryotic signature proteins (ESPs) in the expanded Asgard archaea. Red dots indicate that ESP homologues were detected in at least half of genomes of the lineage while blue dots indicate that ESP homologues were detected in fewer than half of the genomes of the lineage.



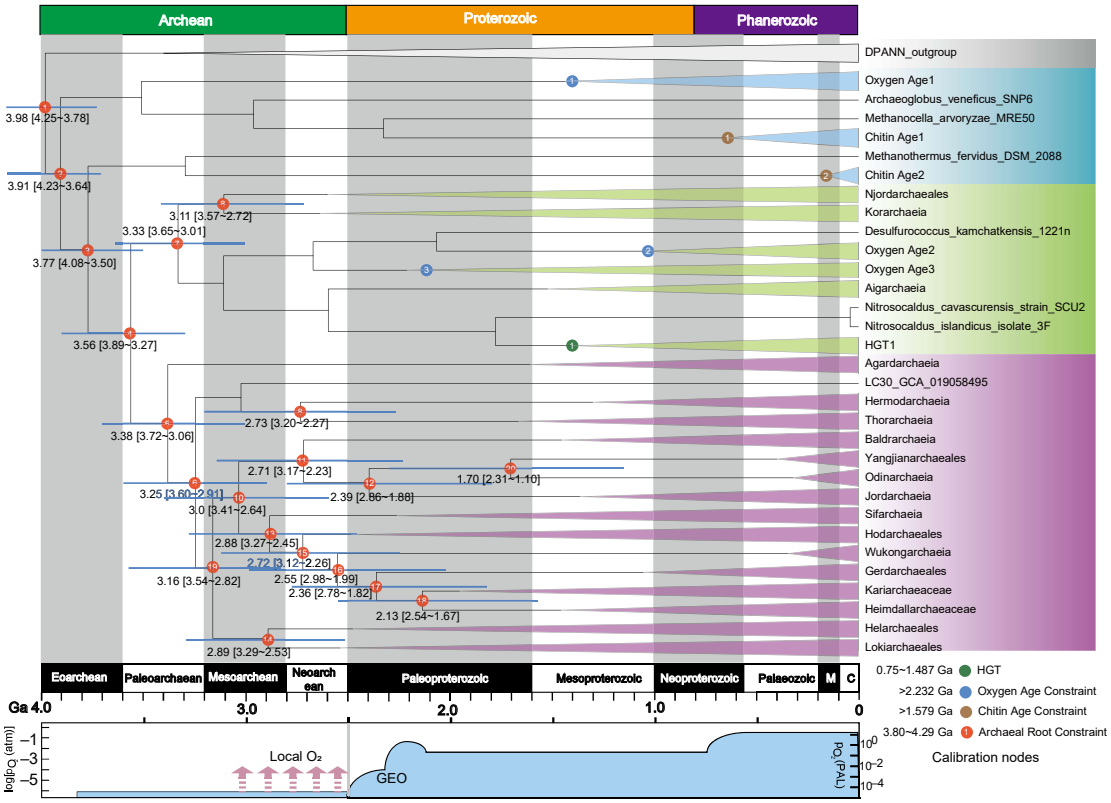
Supplementary Figure 14. Maximum likelihood tree of ε DNA polymerase Pol2, inferred using FastTree2 under LG+CAT model. The scale bar denotes the average expected number of substitutions per site.



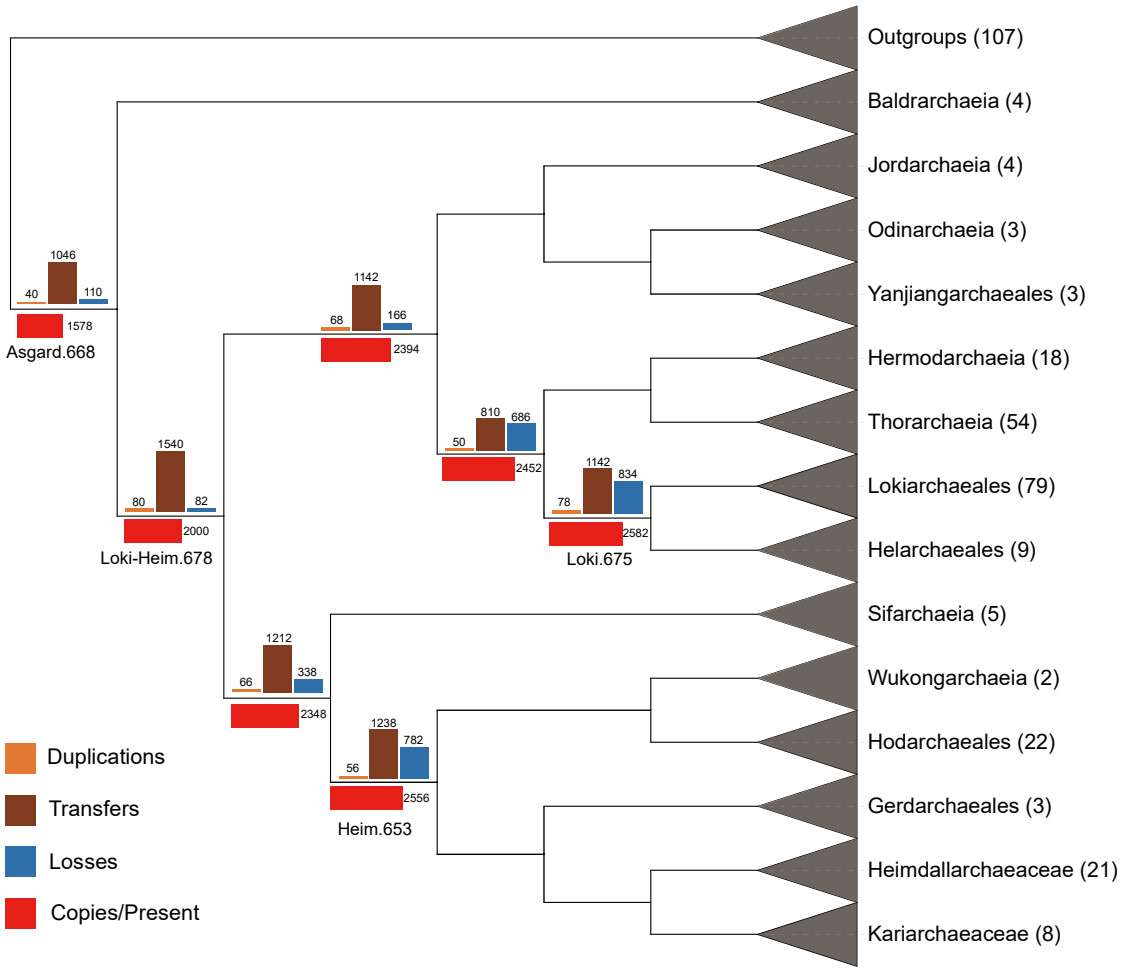
Supplementary Figure 15. Multiple sequence alignment of translation elongation factor 2 (EF2) from Asgard archaea and eukaryotes. The black boxes indicate highly conserved sequence motifs (HxDxxHRG and SPHKHN) in the two loops of bona fide canonical EF2. Asgard/Baldr, Asgardarchaeia/Baldrarchaeia; Jord, Jordarchaeia; Hod, Hodarchaeales.



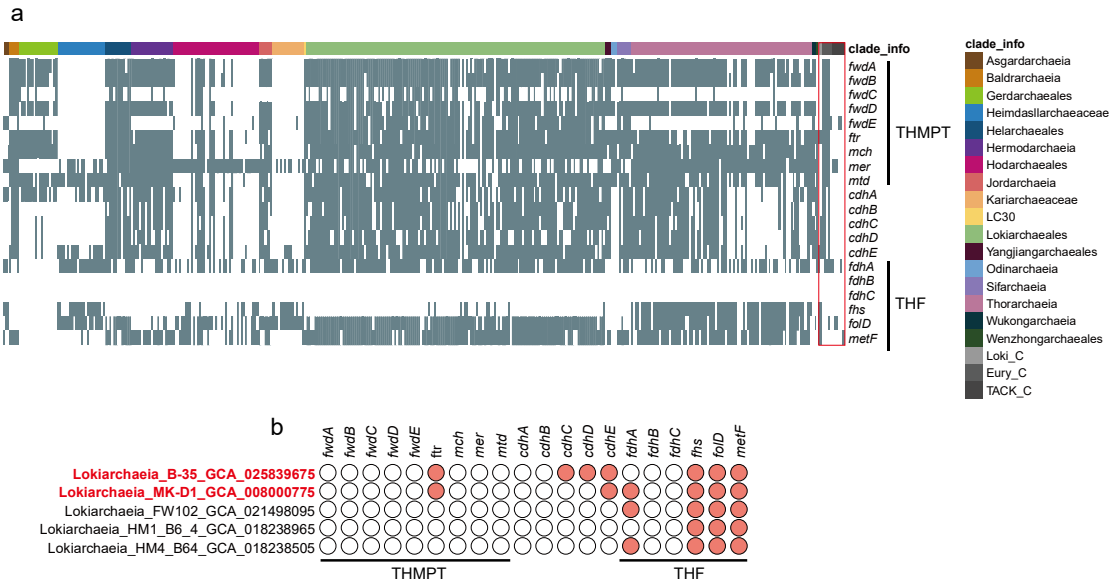
Supplementary Figure 16. Multiple sequence alignment of histones from different archaeal species and human histone H4. The term ‘refs with N tail’ refers to reference sequences containing an N-terminal tail of similar length to that of eukaryotic histones, while ‘refs without N tail’ refers to archaeal histones that lack an N-terminal tail. The histone sequences with an N-terminal tail can be found in Hel, Loki, Thor, Heimdall, and Hod of Asgard archaea. Hel, Helarchaeales; Loki, Lokiarchaeales; Thor, Thorarchaeia; Heimdall, Heimdallarchaeaceae; Hod, Hodarchaeales.



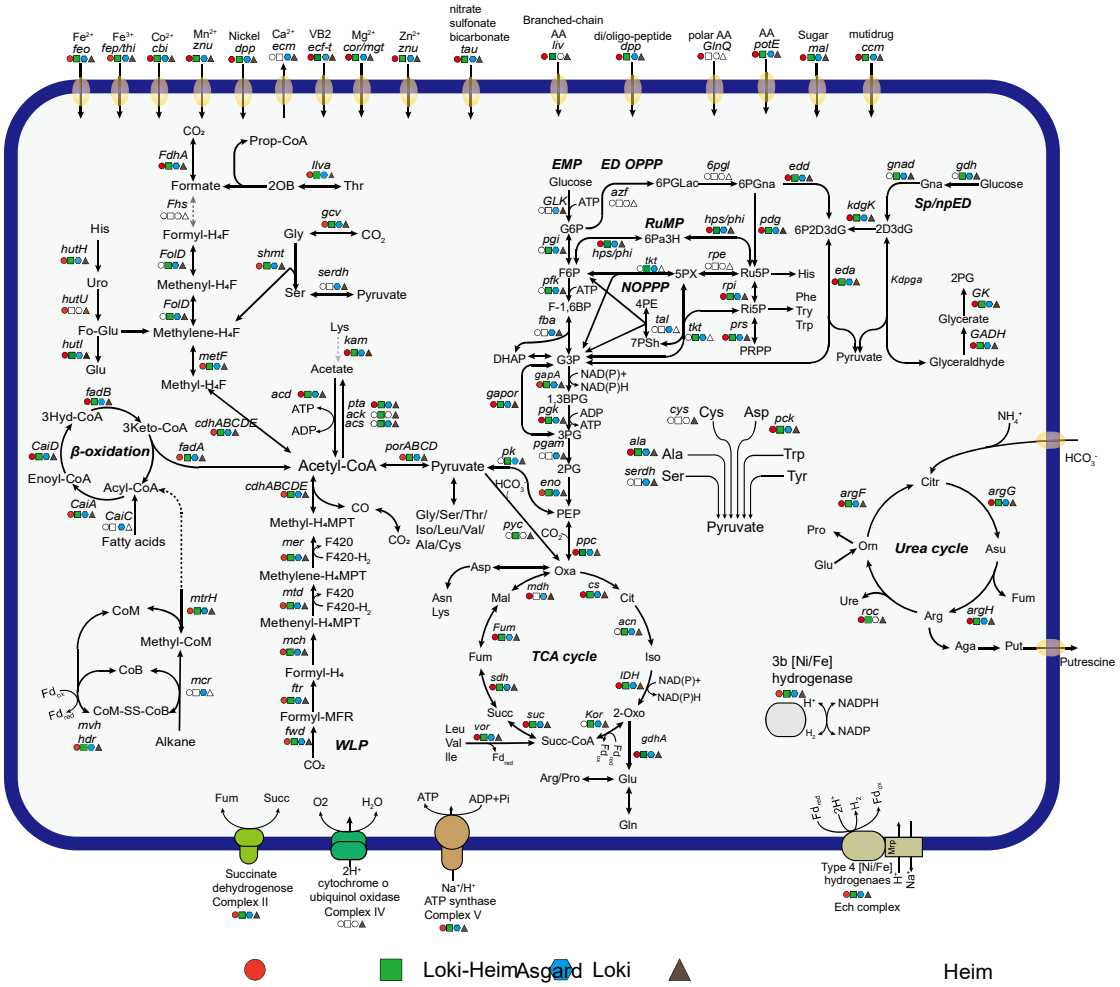
Supplementary Figure 17. Evolutionary history of the Asgard archaea and timing relatedness with major geological events. The phylogenetic tree was constructed using 97 concatenated marker proteins with IQ-TREE under the LG+C60+F+G+PMSF model and dated using MCMCTree. Two DPANN archaea genomes were used as the outgroup. Different color schemes represent archaeal phylum or superphylum: DPANN (gray), Euryarchaea (blue), TACK (green), and Asgard archaea (purple). Calibration nodes are highlighted in colors to represent different types of constraints: archaeal root constraint set at 4.29-3.80 Ga, the chitin age constraint with a minimum of 1.58 Ga, the oxygen age constraint with a maximum of 2.32 Ga, and the horizontal gene transfer constraint from Viridiplantae to Thaumarchaeota set at 1.49-0.75 Ga (Supplementary Table 5). The divergence ages of major nodes are numbered 1 to 20 and labeled with the posterior 95% confidence intervals (CIs), represented by the flanking horizontal blue bars. The bottom panel shows changes of atmospheric oxygen levels on Earth over time[29].



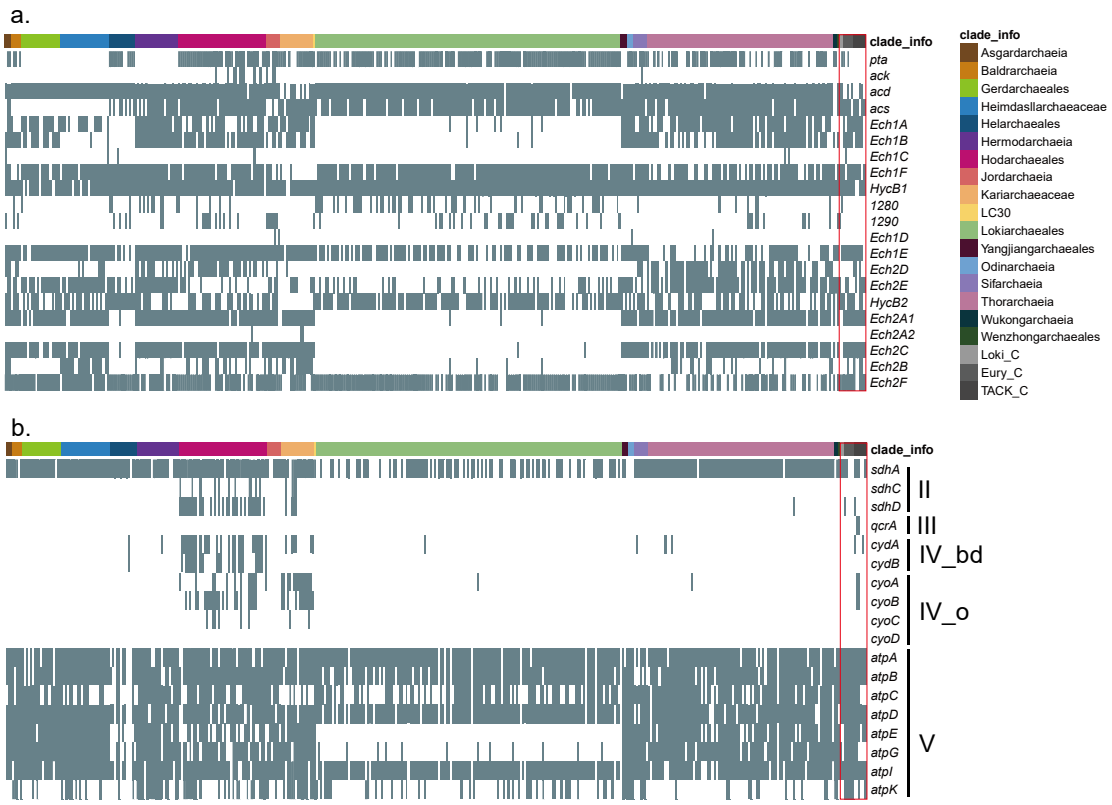
Supplementary Figure 18. Ancestral reconstruction tree. This tree presents genome sizes along with the numbers of gene loss, duplication, and transfer events, inferred using gene-tree species-tree reconciliation method on the S47 dataset for 342 archaeal taxa. The detailed data for the figure is provided in Supplementary Table 8.



Supplementary Figure 19. Distribution of genes associated with Wood-Ljungdahl pathway (WLP) in Asgard archaea. (a) The distribution of genes related to WLP across 411 Asgard archaeal genomes, with the color of the first row representing different clades. Light-blue stripes indicate the presence of genes in a genome. Genes present in cultured archaeal strains are highlighted by a red box. (b) The distribution of genes related to WLP in five genomes from the cluster containing two cultured Asgard strains (B-35 and MK-D1, labelled in red), as shown in phylogenetic tree. Detailed gene names and distribution frequencies are provided in Supplementary table 7. THMPT, tetrahydromethanopterin methyl branch of archaeal WLP; THF, tetrahydrofolate methyl branch of bacterial; Loki_C, cultured Asgard strains; TACK_C, six cultured TACK archaea; Eury_C, five cultured Euryarchaea.



Supplementary Figure 20. Map of metabolic pathways of key ancestral nodes within the Asgard archaea. The red circle denotes the last common ancestor of Asgard archaea. The green square signifies the last common ancestor of Lokiarchaeia and Heimdallarchaeia. The blue hexagon represents the Lokiarchaeia ancestor. The brown triangle marks the Heimdallarchaeia ancestor. The missing genes are indicated by unfilled symbols. Detailed information for genes involved in various metabolic pathways is provided in Supplementary Table 6. EMP, Embden-Meyerhof-Parnas; ED, Entner-Doudoroff; OPPP, oxidative pentose phosphate pathway; NOPPP, nonoxidative pentose phosphate pathway; RuMP, ribulose monophosphate pathway; sp-ED, semi-phosphorylative Entner-Doudoroff pathway; np-ED, Non-phosphorylative Entner-Doudoroff pathway; TCA cycle, citrate cycle; WLP, Wood-Ljungdahl pathway.



Supplementary Figure 21. Distribution of some interesting genes across 411 Asgard archaea. (a) Genes related to acetate metabolism and those encoding Energy-converting hydrogenase (Ech). (b) Genes encoding respiratory chain complexes, The color in the first row represents different clades. Light-blue stripes indicate the presence of genes in a genome. Genes present in cultured archaeal strains are highlighted by a red box. Detailed gene names and distribution frequencies are provided in Supplementary table 7. Loki_C, cultured Asgard strains; TACK_C, six cultured TACK archaea; Eury_C, five cultured Euryarchaea. II, Succinate dehydrogenase; III, Cytochrome c reductase iron-sulfur subunit; IV_bd, Cytochrome bd ubiquinol oxidase complex; IV_o, Cytochrome o ubiquinol oxidase; V, Archaeal-type ATP synthase.

Legends for Supplementary Tables.

Supplementary Table 1. Sampling information.

Supplementary Table 2. Information of Asgard archaeal metagenome-assembled genomes (MAGs), proposed taxonomy.

Supplementary Table 3. Information of reference genomes used for phylogenetic trees in this study. Taxon classification is derived from GTDB r207 or NCBI taxonomy.

Supplementary Table 4. The marker sets used in this study.

Supplementary Table 5. The genomes and calibration nodes used for phylogenetic divergence time estimation.

Supplementary Table 6. Reconciliation frequency of genes in key Asgard ancestors inferred by ALE. The value that is greater than or equal to 0.1 is highlighted in green, indicating the presence of the gene at that node.

Supplementary Table 7. Main metabolic pathways in Asgard archaea and their corresponding gene frequencies. A gene present in multiple copies within a genome is counted only once. The number following each Asgard lineage represents the number of genomes.

Supplementary Table 8. Results of amalgamated likelihood estimation (ALE) analyses. Events of loss, transfer, origination, duplication as well as presence/absence (copies) were counted if they had a raw reconciliation frequency of at least 0.1.

Supplementary Table 9. Distribution of Eukaryotic signature proteins (ESP) in Asgard lineages. The value of '0' indicates that the ESP is not encoded in the clade; '0.5' indicates that no more than 50% of the genomes in the clade encode the ESP; '1' indicates that more than 50% of the genomes in the clade encode the ESP. See supplementary dataset for original protein sequences in Figshare (https://figshare.com/s/6e523322b0b647b91dda).