




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 Anda⁵, Max E. Schön¹, Kiley W. Seitz^{5†††}, Nina Dombrowski^{5††††}, William H. Lewis^{1,3†††††}, Felix Homa³, Jimmy H. Saw^{1††††††}, Jonathan Lombard¹, Takuro Nunoura⁶, Wen-Jun Li⁷, Zheng-Shuang Hua⁸, Lin-Xing Chen⁹, Jillian F. Banfield^{9,10}, Emily St John¹¹, Anna-Louise Reysenbach¹¹, Matthew B. Stott¹², Andreas Schramm¹³, Kasper U. Kjeldsen¹³, Andreas P. Teske¹⁴, Brett J. Baker^{5,15}, Thijs J. G. Ettema^{1,3*}

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

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

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

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

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

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

⁷State 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

⁸Chinese 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

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

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

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

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

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

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

¹⁵Department 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

In the ongoing debates about eukaryogenesis, the series of evolutionary events leading to the emergence of the eukaryotic cell from prokaryotic ancestors, members of the Asgard archaea play a key role as the closest archaeal relatives of eukaryotes. However, the nature and phylogenetic identity of the last common ancestor of Asgard archaea and eukaryotes remain unresolved. Here, we analyze distinct phylogenetic marker datasets of an expanded genomic sampling of Asgard archaea and evaluate competing evolutionary scenarios using state-of-the-art phylogenomic approaches. We find that eukaryotes are placed, with high confidence, as a well-nested clade within Asgard archaea, as a sister lineage to Hodarchaeales, a newly

proposed order within Heimdallarchaea. Using sophisticated gene tree/species tree reconciliation approaches, we show that, in analogy to the evolution of eukaryotic genomes, genome evolution in Asgard archaea involved significantly more gene duplication and fewer gene loss events compared to other archaea. Finally, we infer that the last common ancestor of Asgard archaea likely was a thermophilic chemolithotroph, and that the lineage from which eukaryotes evolved adapted to mesophilic conditions and acquired the genetic potential to support a heterotrophic lifestyle. Our work provides key insights into the prokaryote-to-eukaryote transition and the platform for the emergence of cellular complexity in eukaryotic cells.

Main

Understanding how complex eukaryotic cells emerged from prokaryotic ancestors represents a major challenge in biology^{1,2}. A main point of contention in refining eukaryogenesis scenarios revolves around the exact phylogenetic relationship between Archaea and eukaryotes. The use of phylogenomic approaches with improved models of sequence evolution combined with a much-improved archaeal taxon sampling – progressively unveiled by metagenomics – has recently yielded strong support for the “two-domain” tree of life, in which the eukaryotic clade branches from within Archaea^{3–8}. The discovery of the first Lokiarchaea genome provided additional evidence for the two-domain topology since this lineage was shown to represent, at the time, the closest relative of eukaryotes in phylogenomic analyses⁹. Moreover, Lokiarchaea genomes were found to uniquely contain many genes encoding eukaryotic signature proteins (ESPs) –proteins involved in hallmark complex processes of the eukaryotic cell–, more so than any other prokaryotic lineage. The subsequent identification and analyses of several diverse relatives of Lokiarchaea,

together forming the Asgard archaea superphylum, confirmed that Asgard archaea represented the closest archaeal relatives of eukaryotes^{2,9,10}. Their exact evolutionary relationship to eukaryotes, however, remained unresolved: it has been unclear whether eukaryotes evolved from *within* Asgard archaea, or if they represented their sister-lineage¹⁰. Furthermore, two studies questioned this view of the tree of life altogether, suggesting that Asgard archaea represent a deep-branching Euryarchaeota-related clade^{11,12}, and that, in accordance with the “three-domain” tree, eukaryotes represent a sister group to all Archaea, although this was challenged^{13,14}. A follow-up study that included an expanded taxonomic sampling of Asgard archaeal genome data failed to resolve the phylogenetic position of eukaryotes in the tree of life¹⁵.

Here, we expand the genomic diversity of Asgard archaea by generating 63 novel Asgard metagenomic-assembled genomes (MAGs) from samples from 11 locations around the world. By analyzing the improved genomic sampling of Asgard archaea using state-of-the-art phylogenomics, including recently developed gene tree/species tree reconciliation approaches for ancestral genome content reconstruction, we firmly place eukaryotes nested within the Asgard archaea. By revealing key features regarding the identity, nature and physiology of the last Asgard archaea and eukaryotes common ancestor (LAECA), our results represent important, thus far missing pieces of the elusive eukaryogenesis puzzle.

Expanded Asgard archaea genomic diversity

To increase the genomic diversity of Asgard archaea, we sampled aquatic sediments and hydrothermal deposits from eleven geographically distinct sites (Supplementary Table 1, Supplementary Figure 1). After extraction and sequencing of total environmental DNA, we assembled and binned metagenomic reads into MAGs. Of these MAGs, 63 were found to belong

to the Asgard archaea superphylum, with estimated median completeness and redundancy of 83% and 4.2%, respectively (Supplementary Table 1). To assess the genomic diversity in this dataset, we reconstructed a phylogeny of ribosomal proteins encoded in a conserved 15-ribosomal protein (RP15) gene cluster¹⁶ from these MAGs, and all publicly available Asgard archaea assemblies (retrieved June 29th, 2021; Figure 1). These analyses expand the genomic sampling across previously described major Asgard archaea clades (i.e., Loki-, Thor-, Heimdall-, Odin-, Hel-, Hermod-, Sif-, Jord- and Baldrarchaeia^{9,10,15,17,18}) and recover a previously undescribed clade of high taxonomic rank (Candidatus Asgardarchaeia; see Ext. Data Fig. 1 and Supplementary Information for proposed uniformization of Asgard archaea taxonomic classification that will be adhered to throughout the present manuscript). We observed that the median estimated Asgard archaeal genome size (3.8 Mega basepairs (Mbp)) is considerably larger than those of representative genomes from TACK archaea and Euryarchaeota (median=1.8 Mbp for both) and DPANN archaea (median=1.2 Mbp) (Supplementary Table 1). Among Asgard archaea, Odinararchaeia display the smallest genomes (median=1.4 Mb), while Loki- and Helarchaeales contain the largest (median=4.3 Mbp for both). Unlike other major Asgard archaeal clades, Heimdallarchaeia possess a wide range of genome sizes, spanning from 1.6 to 7.4 Mbp (median=3.5 Mbp). Indeed, this large class contains five clades with diverse features. These include Njordarchaeales (median genome size=2.4 Mbp) followed by Kariarchaeaceae (median genome size=2.7 Mbp), Gerdarchaeales (median genome size=3.4 Mbp), Heimdallarchaeaceae (median genome size=3.7 Mbp), and finally Hodarchaeales (median genome size=5.1 Mbp). The smallest heimdallarchaeial genome corresponds to the only Asgard archaeal MAG recovered from a marine surface water metagenome (Heimdallarchaeota archaeon RS678)¹⁹, in agreement with reduced genome sizes typically observed among prokaryotic plankton of the euphotic zone²⁰.

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}, Helarchaeia¹⁰, for which we here present 36 new genomes, and the
 93 recently proposed Sif-¹⁸, Hermod-¹⁷, Jord-²¹, Wukong-¹⁵ and Baldrarchaeia¹⁵, for most of which
 94 we also identified new near-complete MAGs. Finally, we identified 15 MAGs representing the
 95 recently described Njordarchaeales²² (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 ESPs¹⁰,
 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

the base of the clade formed by Korarchaeota and Njordarchaeales, although with weak support. Hereafter, we focused on disentangling the historically correct phylogenetic signal from noise and artefacts.

Alternative phylogenomic markers

Despite often being used in phylogenomic analyses, ribosomal proteins have been suggested to contribute to phylogenetic artefacts due to inherent compositional sequence biases^{23,24}. Additionally, considering the inconsistency of the obtained placement of eukaryotes compared to previous analyses, the incoherent placement of Njordarchaeales, and the presence of long branches at the base of both of these clades in the RP56 tree, we sought to use an alternative phylogenetic marker set to obtain a stable Asgard archaeal species tree, and to further investigate the phylogenetic position of eukaryotes. We constructed an independent ‘new marker’ dataset comprising 57 proteins of archaeal origin in eukaryotes (NM57 dataset; see Methods). The NM57 proteins are mostly involved in diverse informational, metabolic, and cellular processes, but do not include ribosomal proteins (Supplementary Table 2). Besides being longer, and hence putatively more phylogenetically informative compared to the RP56 markers, the broader functional distribution of NM57 markers is less likely to cause phylogenetic reconstruction artefacts induced by strong co-evolution between proteins – something that is to be expected for functionally and structurally cohesive ribosomal proteins²⁵. Indeed, in case co-evolving protein sequences are compositionally biased, and hence violate evolutionary model assumptions of fixed composition over species, their concatenation is expected to strengthen the artefactual, non-phylogenetic signal and the statistical support for incorrect relationships²⁶. We thus decided to independently evaluate the concatenated NM57 and RP56 marker datasets for downstream

phylogenomic analyses. We observed that ML phylogenomic analyses of the NM57 dataset did not only recover Njordarchaeales as *bona fide* Asgard archaea, they were also placed as the closest relatives of eukaryotes (BS=98%; Supplementary Figure 3), as was proposed in a recent analysis²². To investigate the underlying causes for the contradicting results between the NM57 and RP56 datasets, we first assessed the effect of taxon sampling on phylogenetic reconstructions by removing eukaryotic and/or DPANN and/or Korarchaeota sequences from the alignments, for two main reasons: (1) eukaryotes and DPANN archaea represent long-branching clades potentially inducing long branch attraction (LBA) artefacts; and (2) we wanted to investigate the effects of removing eukaryotes and Korarchaeota, which were the sister lineages of Njordarchaeales in the NM57 and RP56 phylogenetic analyses, respectively. Following this, we recoded the alignments into 4 states (using SR4-recoding²⁷) to ameliorate potential phylogenetic artefacts arising from model misspecification at mutationally saturated or compositionally biased sites^{14,28–30}. Further, with a similar goal, we applied a fast-evolving site removal (FSR) procedure to the concatenated datasets, since fast-evolving sites are often mutationally saturated. We performed phylogenetic analyses of the above-mentioned datasets in both ML and Bayesian Inference (BI) frameworks, under sophisticated evolutionary models that account for sequence heterogeneity in the substitution process across sites (mixture models; Supplementary Table 2).

Phylogenomic analyses of the above-mentioned combinations of taxon sampling, data treatments and phylogenetic frameworks revealed that Njordarchaeales are artefactually attracted to Korarchaeota in RP56 datasets (Supplementary Information). This attraction is likely caused by the high compositional similarity of njord- and korarchaeal RP56 ribosomal protein sequences, which is probably linked to their shared hyperthermophilic lifestyle (Supplementary Figures 4-6). Analyses of RP56 datasets from which Korarchaeota were removed, recovered Njordarchaeales as

an order at the base of or within Heimdallarchaeia (Supplementary Figure 7), consistent with phylogenomic analyses of the NM57 dataset that included Korarchaeota (Supplementary Figure 3). Next, in our efforts to resolve the phylogenetic placement of eukaryotes, we initially performed phylogenomic analyses on variations of the RP56 and NM57 datasets (Supplementary Table 2 and Discussion). However, since compared to the RP56 dataset, the NM57 dataset is larger and less compositionally biased, and is thus expected to have retained a stronger and more congruent phylogenetic signal, we focused the rest of our study on this more reliable dataset.

Eukarya represent a well-nested clade within Heimdallarchaeia

Subsequent phylogenetic analyses of untreated NM57 datasets with various taxon sampling variations recovered eukaryotes as sister-clade to Njordarchaeales in ML analyses (e.g., Supplementary Figure 3, Supplementary Table 2 and Supplementary information). However, ML analyses of the SR4-recoded datasets retrieved a complex phylogenetic signal, as in some cases eukaryotes were placed at the base of all Heimdallarchaeia (including Njordarchaeales) and Wukongarchaeia. This strongly suggests that the previously observed phylogenetic affiliation between Njordarchaeales and eukaryotes could represent an artefact. Furthermore, when both SR4-recoding and FSR treatments were combined, eukaryotes were nested within Heimdallarchaeia, as sister-group to the order Hodarchaeales (Figure 2; Supplementary Figure 8), and this position was supported by ML analyses of NM57 datasets across all taxon selection variations (removing DPANN archaea, and/or Korarchaeota and/or Njordarchaeales). Congruently, the monophyly of eukaryotes and Hodarchaeales was systematically recovered by BI of recoded datasets (both with and without FSR; Figure 2, Supplementary Table 2). In addition, the position of Njordarchaeales shifted during these analyses, moving from a deep position at the base of Heimdallarchaeia and

Wukongarchaeia, to a more nested position forming a clade with Gerdarchaeales and Kari- and Heimdallarchaeaceae (Supplementary Discussion). This shift is observed in both the NM57 and the RP56 datasets analyses when SR4-recoding and FSR was combined (Supplementary Figures 9-10), supporting that Njordarchaeales represent a divergent order-level lineage of Heimdallarchaeia.

In summary, resolving the position of eukaryotes relative to Asgard archaea is anything but trivial (see Supplementary Discussion). In our efforts to extract the true phylogenetic signal, we provide confident support for eukaryotes forming a well-nested clade within the Asgard archaea phylum, consistent with the 2D tree of life scenario. More specifically, we observe that eukaryotes affiliate with the Heimdallarchaeia in analyses in which we systematically reduce phylogenetic artefacts, predominantly converging on a position of eukaryotes as sister to Hodarchaeales, which is also in line with the observed ESP content and genome evolution dynamics (see below).

Informational ESPs in Hodarchaeales

We found that most of the ESPs previously identified in a limited sampling of Asgard archaea^{9,10} are widespread across all phyla included in the present study (Figure 3, Supplementary Table 3). Notably, we observed some exceptions in support of the phylogenetic affiliation between Hodarchaeales and eukaryotes, particularly among ESPs involved in information processing: (1) the ϵ DNA polymerase subunit is only found in Hodarchaeales; (2) ribosomal protein L28e (Rpl28e/Mak16) homologs are unique to Njord- and Hodarchaeales members; (3) many archaea that lack genes coding for the synthesis of diphthamide, a modified histidine residue which is uniquely present in archaeal and eukaryotic elongation factor 2 (EF-2), instead encode a second EF-2 paralog that misses key-residues required for diphthamide modification³¹. Interestingly, we

found that among all Asgard archaea, only MAGs of all sampled Hodarchaeales members encode *dph* genes in addition to a single gene encoding canonical EF-2, which branches at the base of their eukaryotic counterparts in phylogenetic analyses (Supplementary Figure 11; Supplementary Information); (4) While RPL22e and RNA polymerase subunit RPB8 are found in several Asgard archaeal phyla, the only Heimdallarchaeia genomes encoding these genes are members of the Hodarchaeales. Finally, (5) we identified N-terminal histone tails characteristic of eukaryotic histones in all three Hodarchaeales MAGs, as well as in three Njordarchaeales genomes (see Supplementary Information). Altogether, the identification of these key-informational ESPs, in agreement with phylogenomic analyses described above, supports that Hodarchaeales represent the closest archaeal relatives of eukaryotes.

Expanded eukaryotic-like protein translocation repertoire

In our search for putative new ESPs in the expanded Asgard archaeal genomic diversity, we uncovered several additional homologs of proteins associated with the eukaryotic translocon, a protein complex primarily responsible for the post-translational modification of proteins, and subsequent insertion into, or transport across the membrane of the endoplasmic reticulum (ER)³². The eukaryotic translocon is comprised of the core Sec61 protein-conducting channel, and several accessory components, including the oligosaccharyltransferase (OST) and translocon-associated protein (TRAP) complexes (Figure 3b), both of which are involved in the biogenesis of N-glycosylated proteins³³. The eukaryotic TRAP complex is composed of two to four subunits in eukaryotes. Using distant-homology detection methods, we identified homologs from three of these subunits to be broadly distributed across Asgard archaeal genomes, while the fourth one was detected only in a few thorarchaeial MAGs (Figure 3b). The eukaryotic OST complex generally

comprises 6-8 subunits organized into three subcomplexes that are collectively embedded in the ER membrane³⁴ (Figure 3b). Apart from STT3/AglB (OST subcomplex-II), which represents the catalytic subunit and is universally found across all three domains of life, other OST subcomplexes generally do not possess prokaryotic homologs beyond the Ost1/Ribophorin I (OST subcomplex-I) and Ost3/Tusc3 (OST subcomplex-II) subunits previously reported in Asgard archaea¹⁰. Here, we report the identification of Asgard archaeal homologs of all five additional subunits, Ost2/Dad1, Ost4, Ost5/TMEM258, SWP1/Ribophorin II and WBP1/Ost48. While we identified homologs of Ost4 and Ost5 (OST subcomplex-I) in most Asgard archaeal classes, the distribution of Ost2, WBP1, and Swp1, the first subcomplex-III subunits described in prokaryotes to date, was restricted to Heimdallarchaeia, including Njordarchaeales for WBP1, further supporting their monophyly. Our findings indicate that Asgard archaea and, by inference, LAECA, potentially encode relatively complex machineries for the N-linked glycosylation and translocation of proteins (Figure 3b).

Vesicular biogenesis and trafficking proteins

Intracellular vesicular transport represents a key process that emerged during eukaryogenesis. Previous studies have reported that Asgard archaeal genomes encode homologs of eukaryotic proteins comprising various intracellular vesicular trafficking and secretion machineries, including the ESCRT (endosomal sorting complexes required for transport), TRAPP (transport protein particle) and COPII (coat protein complex II) vesicle coatomer protein complexes^{9,10}. Furthermore, as much as 2% of the genes of Asgard archaeal genomes were found to encode small GTPase homologs – a broad family of eukaryotic proteins, encompassing the Ras, Rab, Arf, Rho and Ran subfamilies, that are broadly implicated in budding, transport, docking and fusion of vesicles in

eukaryotic cells^{9,10}. Here, we report the identification of Asgard archaeal homologs of subunits of additional vesicular trafficking complexes (Figure 3, Ext. Data Fig. 2, Supplementary Table 3). Noticeably, we found putative homologs of all four subunits composing eukaryotic adaptor proteins (AP) and coatamer protein (COPI) complexes, which, in eukaryotic cells, are involved in the formation of clathrin-coated pits and vesicles responsible for packaging and sorting cargo for transport through the secretory and endocytic pathways³⁵. Those complexes are composed of two large subunits, belonging to the β - and γ -families, a medium μ -subunit, and a small σ -subunit. We found homologs of all functional domains composing those subunits, albeit sparsely distributed (Ext. Data Fig. 2, Supplementary Information). Additionally, we found homologs of several protein complexes involved in eukaryotic endosomal sorting such as the retromer, the HOPS/CORVET and the GARP complexes (Figure 3, red shading). Retromer is a coat-like complex associated with endosome-to-Golgi retrograde traffic³⁶ and we detected four of its five subunits in Asgard MAGs. One of these subunits is Vps5-BAR, which in Thorarchaeia is often fused to Vps28, a subunit of the ESCRT-I subcomplex, suggesting a functional link between BAR domain proteins and the thorarchaeial ESCRT complex. The GARP (Golgi-associated retrograde protein) complex is a multisubunit tethering complex located at the trans-Golgi network in eukaryotic cells, where it also functions to tether retrograde transport vesicles derived from endosomes^{37,38}, similarly to the retromer. GARP comprises four subunits, three of which we could detect in Asgard archaeal genomes, with a sparse and punctuated distribution. Functioning in opposite direction from the retromer and GARP complexes are the CORVET (Class C core vacuole/endosome tethering) and HOPS (Homotypic fusion and protein sorting) complexes³⁹. Endosomal fusion and autophagy in eukaryotic cells depend on them and they share four core

subunits⁴⁰, three of which can be found in Asgard archaea, in addition to one of the HOPS specific subunits⁴¹.

Finally, while numerous components of the ESCRT-I, II and III systems have been previously detected in Asgard archaea^{9,10,42}, we report here the identification of Asgard homologs for the ESCRT-III regulators Vfa1, Vta1, Ist1, and Bro1.

Ancestral Asgard archaea genome reconstruction

The analysis of Asgard archaeal genome data obtained through metagenomics, combined with the insights derived from cytological observations of the first two cultured Asgard archaea ‘*Candidatus* Prometheoarchaeum syntrophicum’⁴³ and ‘*Candidatus* Lokiarchaeum ossiferum’⁴⁴, have generated new hypotheses about the nature of the archaeal ancestor of eukaryotes^{43,45–48}. However, these theories are mostly based on a limited number of features displayed by a single, or a few Asgard archaeal lineages. While informative, features of present-day Asgard archaea do not necessarily resemble those of LAECA, as these are potentially separated by over 2 Gya of evolution⁴⁹. Furthermore, Asgard archaeal phyla display a highly variable genome content with respect to ESPs and predicted metabolic features^{43,46,48,50,51}, suggesting a complex evolutionary history of those traits. In light of these considerations, we inferred ancestral features of LAECA by using an ML evolutionary framework. We employed a recently developed probabilistic gene-tree species-tree reconciliation approach^{52,53} in combination with the extended taxonomic sampling of Asgard archaeal genomes to reconstruct the evolutionary history of homologous gene families and ancestral gene content across the Asgard archaeal species tree. For this, we inferred ML phylogenetic trees of all 17,200 protein families encoded across 181 archaeal genomes, including representatives from Asgard and TACK archaea, and Euryarchaeota clades. Importantly,

as missing genes and potential contaminations in MAGs will be regarded as recent gene loss and gain events in our ancestral reconstruction analyses, the use of incomplete MAGs with low contamination levels is unlikely to have a major impact on the inferred gene content of the deep archaeal ancestors that were reconstructed in the present study (also see Supplementary Information).

We first compared the distributions of estimated ancestral proteome sizes, and numbers of inferred gene duplications, losses and gains (i.e., horizontal gene transfers and originations) in all archaeal ancestral nodes (Supplementary Figure 12). Intriguingly, we observed that Heimdall- (and particularly the ancestor of Hodarchaeales) and Lokiarchaeia ancestors display significantly higher gene duplication rates compared to TACK and Euryarchaeota ancestors (Figure 4a). In addition, we found that most Asgard archaeal ancestors displayed gene loss rates comparable to other archaea, with the exception of Thorarchaeia, Lokiarchaeales and Jordarchaeia, which showed significantly lower loss rates. In agreement with the observed evolutionary genome dynamics, we found that predicted proteome sizes of most Asgard archaea ancestors are significantly larger than other archaeal ancestors ($P < 0.001$), with Lokiarchaeia ancestors displaying the largest estimated proteome size (Supplementary Figure 13). Similarly, the Hodarchaeales ancestor had an estimated proteome size of 4,053 proteins, versus 3,134 for the last Asgard archaea common ancestor (LAsCA), reflecting the high duplication and low loss rates in that clade. The streamlined genome content of the Odinararchaeia ancestor represents an exception to the general trend of genome expansion across Asgard archaea, and possibly reflects an adaptation to high temperatures⁵⁴.

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 LAECA 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

(NarGHIJ), indicating that nitrate might have been used as a terminal electron acceptor to perform anaerobic respiration. As such, the last Hodarchaeales common ancestor likely generated ATP using an electron transport chain where electrons from NADH and succinate were transferred through a series of membrane-associated complexes with quinones and cupredoxins as electron carriers to ultimately reduce nitrate⁵⁵.

As indicated above, a significant fraction of the currently sampled Asgard archaea diversity originates from geothermal or hydrothermal environments. Indeed, using an algorithm based on genome-derived features⁵⁶, we confirmed that (most) Njordarchaeales, Baldr- and Jordarchaeia are hyperthermophiles, Odinararchaeia are thermophiles, and Loki- and Thorarchaeia are mesophiles (Figure 4c, Supplementary Table 5). While Heimdallarchaeia seem to contain both meso- and thermophiles, we infer a mesophilic physiology for Hodarchaeales, obtaining the lowest predicted optimal growth temperatures among all Asgard archaea (median=36.7 °C). Asgard archaeal hyperthermophiles contain reverse gyrase, a topoisomerase that is typically encoded by hyperthermophilic prokaryotes⁵⁷. We infer that a reverse gyrase was possibly present in LAsCA and that it was subsequently lost in all heimdallarchaeial orders except for Njordarchaeales. This observation would be compatible with a scenario in which Asgard archaea have a hyperthermophilic ancestry, but in which eukaryotes evolved from an Asgard archaea lineage that had adapted to mesophilic growth temperatures.

Discussion

Beyond genomic exploration, several studies have started to unveil important physiological, cytological and ecological aspects of Asgard archaea^{43,58–60}. Yet, while such insights are certainly relevant, the cellular and physiological characteristics of present-day Asgard archaea will almost

certainly not resemble those of LAECA. Therefore, inferences about the identity and nature of LAECA and the process of eukaryogenesis should be made within an evolutionary context. We used an evolutionary framework to analyze an expanded Asgard archaeal genomic diversity, comprising 11 clades of high taxonomic rank. Using comprehensive phylogenomic analyses involving the evaluation of distinct marker protein datasets and systematic assessment of suspected phylogenetic artefacts and state-of-the-art models of evolution, we identified Hodarchaeales, a class-level clade within the Heimdallarchaeia, as the closest relatives of eukaryotes. Evidently, phylogenomic analyses aiming to pinpoint the phylogenetic position of eukaryotes in the tree of life are extremely challenging, and our results stress the importance of testing for possible sources of bias affecting phylogenomic reconstructions, as was recently reviewed⁶¹. The implementation of a probabilistic gene tree/species tree reconciliation approach allowed us to infer the evolutionary dynamics and ancestral content across the archaeal species tree providing several new insights into the Asgard archaeal roots of eukaryotes. Altogether, our results reveal a picture in which the Asgard archaeal ancestor of eukaryotes had, compared to other archaea, a relatively large genome, resulting mainly from more numerous gene duplication and fewer gene loss events. It is tempting to speculate that the elevated gene duplication rates observed in our analyses represent an ancestral feature of LAECA, and that it remained the predominant modus of genome evolution during the early stages of eukaryogenesis. We also inferred that the duplicated gene content of LAECA included several protein families involved in cytoskeletal and membrane-trafficking functions, including among others actin homologs, ESCRT complex subunits and small GTPase homologs. Our findings complement those of another study⁶² reporting that eukaryotic proteins with an Asgard archaeal provenance, as opposed to those inherited from the mitochondrial symbiont,

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 pathways³⁵. 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 no⁴³, or only limited⁴⁴ 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

present-day Asgard archaeal lineages do not necessarily resemble those of LAECA. Furthermore, given the disparity of the distribution patterns of membrane trafficking homologs in Asgard archaea, it will be crucial to isolate representatives of phyla other than Lokiarchaeia and study their cell biological features and potential for endomembrane biogenesis. Of particular interest would be members of the Heimdallarchaeia, and specifically Hodarchaeales, as the currently identified closest relatives of eukaryotes, as well as Thorarchaeia lineages, which seem to generally contain a particularly rich suite of homologs of eukaryotic membrane trafficking proteins.

By phylogenetically placing eukaryotes as a firmly-nested clade within the presently identified Asgard archaeal diversity, and by inferring ancestral genomic content across the Asgard archaea, our work provides insights into the identity and nature of the Asgard archaeal ancestor of eukaryotes, guiding future studies aiming to uncover new pieces of the elusive eukaryogenesis puzzle.

References

1. López-García, P. & Moreira, D. Open Questions on the Origin of Eukaryotes. *Trends Ecol. Evol.* **30**, 697–708 (2015).
2. Eme, L., Spang, A., Lombard, J., Stairs, C. W. C. W. & Ettema, T. J. G. T. J. G. Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **15**, nrmicro–2017 (2017).
3. Guy, L. & Ettema, T. J. G. The archaeal ‘TACK’ superphylum and the origin of eukaryotes. *Trends Microbiol.* **19**, 580–587 (2011).
4. Kelly, S., Wickstead, B. & Gull, K. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc. Biol. Sci.* **278**, 1009–1018 (2011).

5. Williams, T. a., Foster, P. G., Cox, C. J. & Embley, T. M. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* **504**, 231–236 (2013).
6. Lasek-Nesselquist, E. & Gogarten, J. P. The effects of model choice and mitigating bias on the ribosomal tree of life. *Mol. Phylogenet. Evol.* (2013) doi:10.1016/j.ympev.2013.05.006.
7. Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences* 201420858 (2015).
8. Cox, C. J., Foster, P. G., Hirt, R. P., Harris, S. R. & Embley, T. M. The archaeobacterial origin of eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 20356–20361 (2008).
9. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
10. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
11. Cunha, V. D. *et al.* Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLOS Genetics* vol. 13 e1006810 Preprint at <https://doi.org/10.1371/journal.pgen.1006810> (2017).
12. Da Cunha, V., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the debate about the universal tree of life topology. *PLoS genetics* vol. 14 e1007215 (2018).
13. Spang, A. *et al.* Asgard archaea are the closest prokaryotic relatives of eukaryotes. *PLoS Genet.* **14**, (2018).
14. Williams, T. A., Cox, C. J., Foster, P. G., Szöllösi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).
15. Liu, Y. *et al.* Expanded diversity of Asgard archaea and their relationships with eukaryotes. *Nature* **593**, 553–557 (2021).
16. Hug, L. A. *et al.* Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**, 22 (2013).
17. Zhang, J.-W. *et al.* Newly discovered Asgard archaea Hermodarchaeota potentially degrade alkanes

- and aromatics via alkyl/benzyl-succinate synthase and benzoyl-CoA pathway. *ISME J.* (2021)
doi:10.1038/s41396-020-00890-x.
18. Farag, I. F., Zhao, R. & Biddle, J. F. ‘Sifarchaeota’ a novel Asgard phylum from Costa Rica
sediment capable of polysaccharide degradation and anaerobic methylotrophy. *Appl. Environ.*
Microbiol. (2021).
19. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-
assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).
20. Swan, B. K. *et al.* Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in
the surface ocean. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11463–11468 (2013).
21. Sun, J. *et al.* Recoding of stop codons expands the metabolic potential of two novel Asgardarchaeota
lineages. *ISME Communications* **1**, 1–14 (2021).
22. Xie, R. *et al.* Expanding Asgard members in the domain of Archaea sheds new light on the origin of
eukaryotes. *Sci. China Life Sci.* **65**, 818–829 (2022).
23. Ribosomal proteins: Toward a next generation standard for prokaryotic systematics? *Mol.*
Phylogenet. Evol. **75**, 103–117 (2014).
24. Petitjean, C., Deschamps, P., López-García, P. & Moreira, D. Rooting the Domain Archaea by
Phylogenomic Analysis Supports the Foundation of the New Kingdom Proteoarchaeota. *Genome*
Biol. Evol. **7**, 191–204 (2014).
25. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. Origins of coevolution between
residues distant in protein 3D structures. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9122–9127 (2017).
26. Foster, P. G. & Hickey, D. A. Compositional bias may affect both DNA-based and protein-based
phylogenetic reconstructions. *J. Mol. Evol.* **48**, (1999).
27. Susko, E. & Roger, A. J. On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol.*
Evol. **24**, 2139–2150 (2007).
28. Brown, M. W. *et al.* Phylogenomics demonstrates that breviate flagellates are related to opisthokonts
and apusomonads. *Proceedings of the Royal Society B: Biological Sciences* **280**, 20131755–

20131755 (2013).

29. Jeffroy, O., Brinkmann, H., Delsuc, F. & Philippe, H. Phylogenomics: the beginning of incongruence? *Trends in Genetics* vol. 22 225–231 Preprint at <https://doi.org/10.1016/j.tig.2006.02.003> (2006).
30. Viklund, J., Ettema, T. J. G. & Andersson, S. G. E. Independent Genome Reduction and Phylogenetic Reclassification of the Oceanic SAR11 Clade. *Molecular Biology and Evolution* vol. 29 599–615 Preprint at <https://doi.org/10.1093/molbev/msr203> (2012).
31. Narowe, 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).
32. Wang, L. & Dobberstein, B. Oligomeric complexes involved in translocation of proteins across the membrane of the endoplasmic reticulum. *FEBS Lett.* **457**, 316–322 (1999).
33. Pfeffer, S. *et al.* Dissecting the molecular organization of the translocon-associated protein complex. *Nat. Commun.* **8**, 14516 (2017).
34. Bai, L., Wang, T., Zhao, G., Kovach, A. & Li, H. The atomic structure of a eukaryotic oligosaccharyltransferase complex. *Nature* **555**, 328–333 (2018).
35. Rout, M. P. & Field, M. C. The Evolution of Organellar Coat Complexes and Organization of the Eukaryotic Cell. *Annu. Rev. Biochem.* **86**, 637–657 (2017).
36. Seaman, M. N. J. The retromer complex - endosomal protein recycling and beyond. *J. Cell Sci.* **125**, 4693–4702 (2012).
37. Liewen, H. *et al.* Characterization of the human GARP (Golgi associated retrograde protein) complex. *Exp. Cell Res.* **306**, 24–34 (2005).
38. Pérez-Victoria, F. J. *et al.* Structural basis for the wobbler mouse neurodegenerative disorder caused by mutation in the Vps54 subunit of the GARP complex. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12860–12865 (2010).
39. Villaseñor, R., Kalaidzidis, Y. & Zerial, M. Signal processing by the endosomal system. *Curr. Opin. Cell Biol.* **39**, 53–60 (2016).

40. Graham, S. C. *et al.* Structural basis of Vps33A recruitment to the human HOPS complex by Vps16. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 13345–13350 (2013).
41. Jiang, P. *et al.* The HOPS complex mediates autophagosome–lysosome fusion through interaction with syntaxin 17. *Molecular Biology of the Cell* vol. 25 1327–1337 Preprint at <https://doi.org/10.1091/mbc.e13-08-0447> (2014).
42. Hatano, T. *et al.* Asgard archaea shed light on the evolutionary origins of the eukaryotic ubiquitin-ESCRT machinery. *Nat. Commun.* **13**, 3398 (2022).
43. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
44. Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* Preprint at <https://doi.org/10.1038/s41586-022-05550-y> (2022).
45. Sousa, F. L., Neukirchen, S., Allen, J. F., Lane, N. & Martin, W. F. Lokiarchaeon is hydrogen dependent. *Nat Microbiol* **1**, 16034 (2016).
46. Bulzu, P.-A. *et al.* Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nature Microbiology* vol. 4 1129–1137 Preprint at <https://doi.org/10.1038/s41564-019-0404-y> (2019).
47. López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).
48. Spang, A. *et al.* Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat Microbiol* **4**, 1138–1148 (2019).
49. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life’s early evolution and eukaryote origin. *Nat Ecol Evol* **2**, 1556–1562 (2018).
50. Seitz, K. W. *et al.* Asgard archaea capable of anaerobic hydrocarbon cycling. *Nat. Commun.* **10**, 1822 (2019).
51. Liu, Y. *et al.* Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. *ISME J.* **12**, 1021–1031 (2018).
52. Szöllösi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the

- space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).
53. Szöllösi, G. J., Davín, A. A., Tannier, E., Daubin, V. & Boussau, B. Genome-scale phylogenetic analysis finds extensive gene transfer among fungi. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140335 (2015).
54. Sabath, N., Ferrada, E., Barve, A. & Wagner, A. Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. *Genome Biol. Evol.* **5**, 966–977 (2013).
55. Savelieff, M. G. *et al.* Experimental evidence for a link among cupredoxins: red, blue, and purple copper transformations in nitrous oxide reductase. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7919–7924 (2008).
56. Sauer, D. B. & Wang, D.-N. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics* Preprint at <https://doi.org/10.1093/bioinformatics/btz059> (2019).
57. Pavel Lulchev, D. K. Reverse gyrase—recent advances and current mechanistic understanding of positive DNA supercoiling. *Nucleic Acids Res.* **42**, 8200 (2014).
58. Akıl, C. & Robinson, R. C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).
59. Orsi, W. D. *et al.* Metabolic activity analyses demonstrate that Lokiarchaeon exhibits homoacetogenesis in sulfidic marine sediments. *Nat Microbiol* **5**, 248–255 (2020).
60. Neveu, E., Khalifeh, D., Salamin, N. & Fasshauer, D. Prototypic SNARE Proteins Are Encoded in the Genomes of Heimdallarchaeota, Potentially Bridging the Gap between the Prokaryotes and Eukaryotes. *Curr. Biol.* **30**, 2468–2480.e5 (2020).
61. Williams, T. A. *et al.* Inferring the deep past from molecular data. *Genome Biol. Evol.* (2021) doi:10.1093/gbe/evab067.
62. Vosseberg, J. *et al.* Timing the origin of eukaryotic cellular complexity with ancient duplications. *Nat Ecol Evol* **5**, 92–100 (2021).

63. Hua, Z.-S. *et al.* Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. *Nat. Commun.* **9**, 2832 (2018).
64. Chen, L.-X. *et al.* Candidate Phyla Radiation Roizmanbacteria From Hot Springs Have Novel and Unexpectedly Abundant CRISPR-Cas Systems. *Front. Microbiol.* **10**, 928 (2019).
65. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
66. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).
67. Dick, G. J. *et al.* Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* **10**, R85 (2009).
68. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
69. Flores, G. E. *et al.* Inter-field variability in the microbial communities of hydrothermal vent deposits from a back-arc basin. *Geobiology* **10**, 333–346 (2012).
70. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
71. Crusoe, M. R. *et al.* The khmer software package: enabling efficient nucleotide sequence analysis. *F1000Res.* **4**, 900 (2015).
72. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* vol. 31 1674–1676 Preprint at <https://doi.org/10.1093/bioinformatics/btv033> (2015).
73. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
74. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
75. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079

(2009).

76. 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).
77. Seitz, K. W., Lazar, C. S., Hinrichs, K.-U., Teske, A. P. & Baker, B. J. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. *ISME J.* **10**, 1696–1705 (2016).
78. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
79. Ounit, R., Wanamaker, S., Close, T. J. & Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* **16**, 236 (2015).
80. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
81. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* vol. 28 1420–1428 Preprint at <https://doi.org/10.1093/bioinformatics/bts174> (2012).
82. Hugoson, E., Lam, W. T. & Guy, L. miComplete: weighted quality evaluation of assembled microbial genomes. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz664.
83. Karst, S. M., Kirkegaard, R. H. & Albertsen, M. mmgenome: a toolbox for reproducible genome extraction from metagenomes. Preprint at <https://doi.org/10.1101/059121>.
84. Albertsen, M. *et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **31**, 533–538 (2013).
85. Dombrowski, N., Teske, A. P. & Baker, B. J. Expansive microbial metabolic versatility and biodiversity in dynamic Guaymas Basin hydrothermal sediments. *Nat. Commun.* **9**, 4999 (2018).
86. Huang, J.-M., Baker, B. J., Li, J.-T. & Wang, Y. New Microbial Lineages Capable of Carbon Fixation and Nutrient Cycling in Deep-Sea Sediments of the Northern South China Sea. *Appl.*

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

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

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

88. Nunoura, T., Inagaki, F., Delwiche, M. E., Colwell, F. S. & Takai, K. Subseafloor microbial communities in methane hydrate-bearing sediment at two distinct locations (ODP Leg204) in the cascadia margin. *Microbes Environ.* **23**, 317–325 (2008).

89. Inagaki, F. *et al.* Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2815–2820 (2006).

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

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

91. 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).

92. Parks, D. H. *et al.* Author Correction: A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1098 (2020).

93. Romalde, J. L., Balboa, S. & Ventosa, A. *Microbial Taxonomy, Phylogeny and Biodiversity*. (Frontiers Media SA, 2019).

94. Wang, H.-C., Minh, B. Q., Susko, E. & Roger, A. J. Modeling Site Heterogeneity with Posterior Mean Site Frequency Profiles Accelerates Accurate Phylogenomic Estimation. *Syst. Biol.* **67**, 216–235 (2018).

95. Lagkouvardos, I. *et al.* IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Sci. Rep.* **6**, 33721 (2016).

96. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* vol. 4 1686 Preprint at <https://doi.org/10.21105/joss.01686> (2019).

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

98. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences. in *Gene Prediction: Methods and Protocols* (ed. Kollmar, M.) 1–14 (Springer New York, 2019).
99. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
100. Miele, V., Penel, S. & Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinformatics* **12**, 116 (2011).
101. Miele, V. *et al.* High-quality sequence clustering guided by network topology and multiple alignment likelihood. *Bioinformatics* **28**, 1078–1085 (2012).
102. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).
103. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9**, 173–175 (2011).
104. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–93 (2016).
105. Makarova, K. S., Wolf, Y. I. & Koonin, E. V. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales. *Life* **5**, 818–840 (2015).
106. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* gkt1223 (2013).
107. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
108. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428**, 726–731 (2016).
109. Søndergaard, D., Pedersen, C. N. S. & Greening, C. HydDB: A web tool for hydrogenase classification and analysis. *Sci. Rep.* **6**, 34212 (2016).
110. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

111. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* vol. 10 845–858 Preprint at <https://doi.org/10.1038/nprot.2015.053> (2015).
112. Petitjean, C., Deschamps, P., López-García, 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).
113. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
114. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
115. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, (2010).
116. Camacho, C. *et al.* BLAST : architecture and applications. *BMC Bioinformatics* vol. 10 421 Preprint at <https://doi.org/10.1186/1471-2105-10-421> (2009).
117. Huerta-Cepas, J. *et al.* Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
118. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
119. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
120. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
121. Guy, L., Roat Kultima, J. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* vol. 26 2334–2335 Preprint at

- https://doi.org/10.1093/bioinformatics/btq413 (2010).
122. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
123. Martijn, J. *et al.* Hikarchaeia demonstrate an intermediate stage in the methanogen-to-halophile transition. *Nature Communications* vol. 11 Preprint at <https://doi.org/10.1038/s41467-020-19200-2> (2020).
124. Huang, W.-C. *et al.* Comparative genomic analysis reveals metabolic flexibility of Woesearchaeota. *Nat. Commun.* **12**, 5281 (2021).
125. Martijn, J., Vosseberg, J., Guy, L., Offre, P. & Ettema, T. J. G. Phylogenetic affiliation of mitochondria with Alpha-II and Rickettsiales is an artefact. *Nature Ecology & Evolution* vol. 6 1829–1831 Preprint at <https://doi.org/10.1038/s41559-022-01871-3> (2022).
126. Dharamshi, J. E. *et al.* Gene gain facilitated endosymbiotic evolution of Chlamydiae. *Nat Microbiol* **8**, 40–54 (2023).
127. Kim, E. *et al.* Implication of mouse Vps26b–Vps29–Vps35 retromer complex in sortilin trafficking. *Biochem. Biophys. Res. Commun.* **403**, 167–171 (2010).
128. Suzuki, S. W., Chuang, Y.-S., Li, M., Seaman, M. N. J. & Emr, S. D. A bipartite sorting signal ensures specificity of retromer complex in membrane protein recycling. *J. Cell Biol.* **218**, 2876–2886 (2019).
129. Balderhaar, H. J. K. & Ungermann, C. CORVET and HOPS tethering complexes - coordinators of endosome and lysosome fusion. *J. Cell Sci.* **126**, 1307–1316 (2013).

Figures

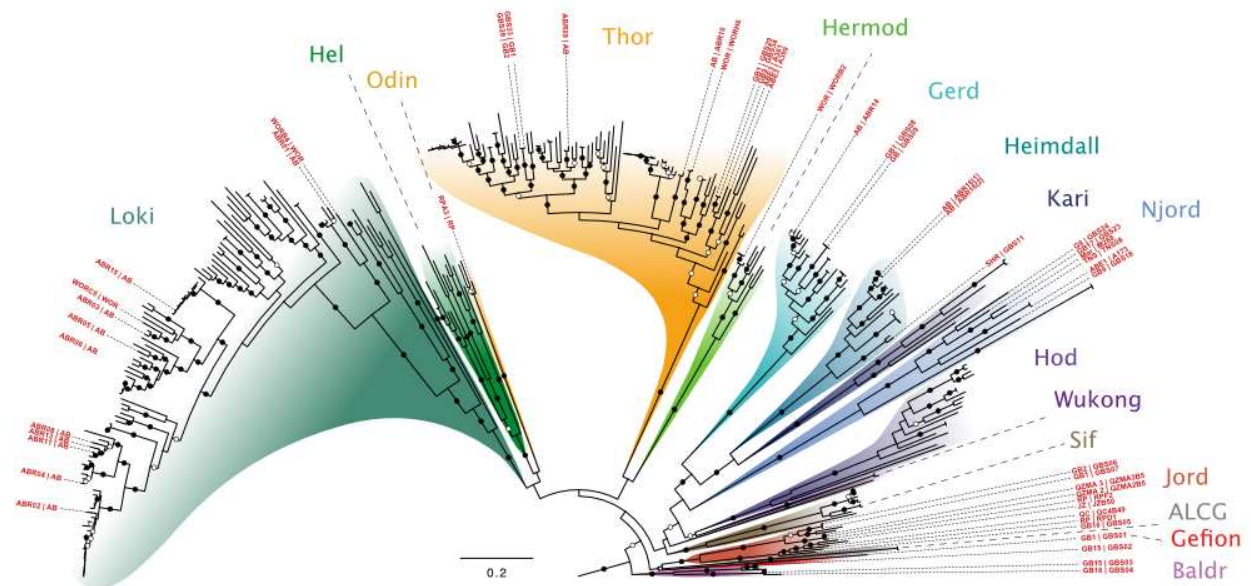


Figure 1. Phylogenomic analysis of 15 concatenated ribosomal proteins expands Asgard archaea diversity. Maximum-likelihood tree (IQ-TREE, WAG+C60+R4+F+PMSF model) of concatenated protein sequences from at least five genes, encoded on a single contig, of a 15 conserved ribosomal protein (RP15) gene cluster 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

(New Zealand); RS: Red sea; SHR: South Hydrate Ridge; TNS: Taketomi Island (Japan); WOR: White Oak River (USA).

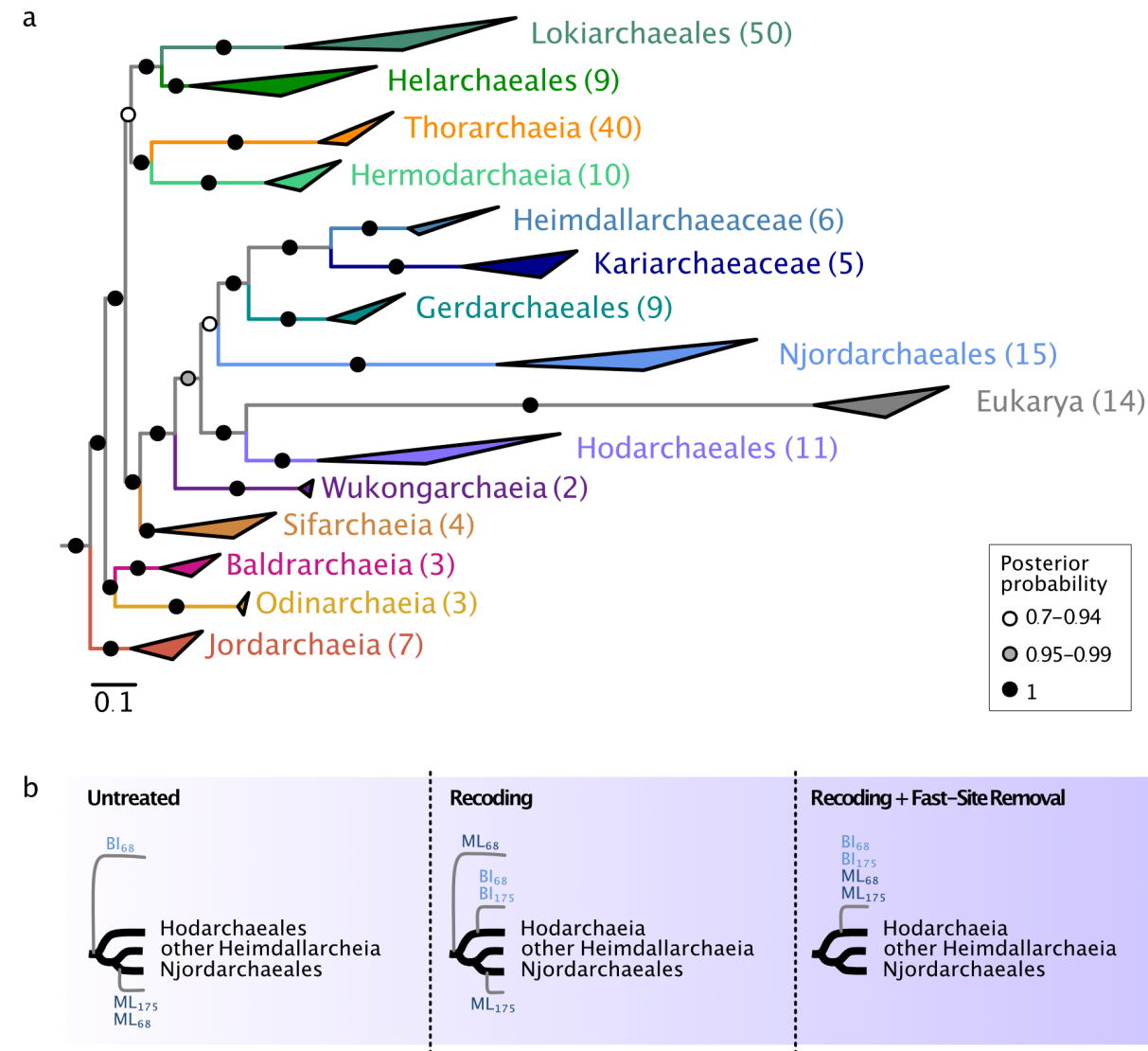


Figure 2. Phylogenomic analyses based on 57 concatenated non-ribosomal proteins support the emergence of eukaryotes as sister to Hodarchaeales. a. Bayesian inference (BI) based on 278 archaeal taxa, using Euryarchaeota and TACK archaea as outgroup (not shown) (NM57-nDK_sr4 alignment, 15,733 amino acid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

analyzed using the CAT+GTR model (4 chains, ~25,000 generations). b. Schematic representation of the shift in the position of eukaryotes (grey branches) in ML and BI analyses of this dataset under different treatments. Untreated: unprocessed dataset; Recoding: SR4-recoded dataset; Recoding+Fast-Site Removal: Fast-site removal combined with SR4-recoding (the topology most often recovered after removing 10% to 50% fastest-evolving sites, in steps of 10%, is shown). 175 and 68 refer to phylogenomic datasets containing 175 and 68 Asgard archaea, respectively. Note that BI was not performed for the 175 untreated dataset due to computational limitations (for detailed overview of phylogenomic analyses, see Supplementary Table 3). Scale bar denotes the average expected number of substitutions per site.

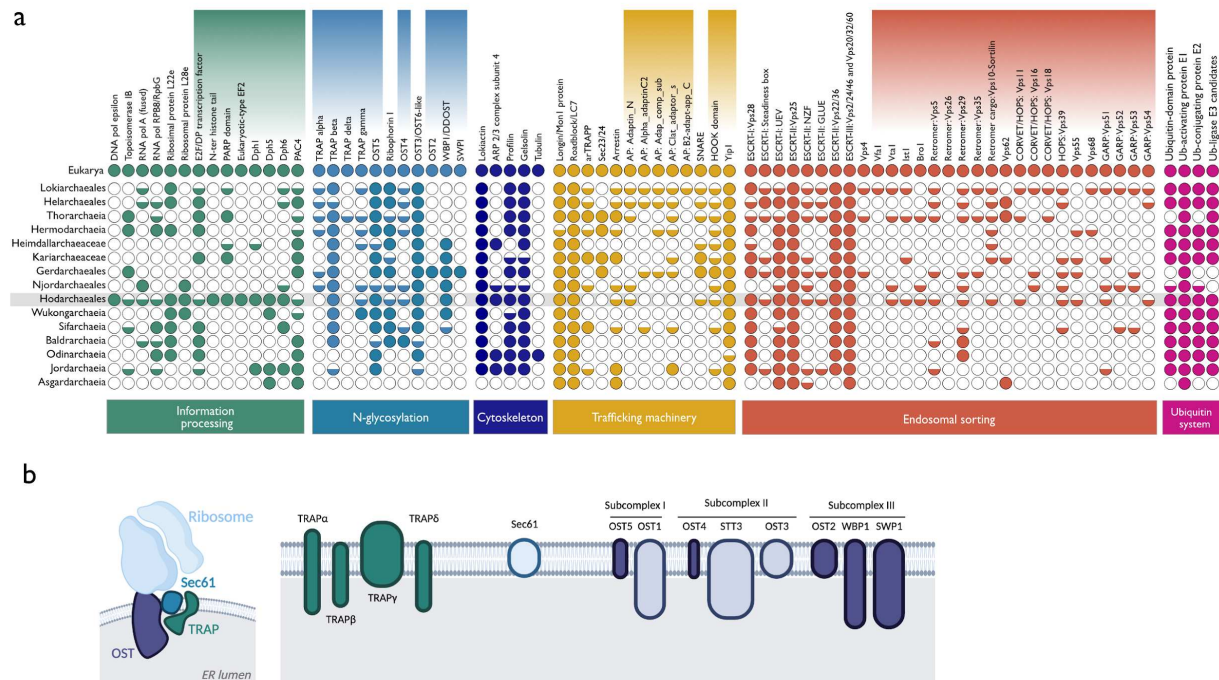


Figure 3. Eukaryotic signature proteins in Asgard archaea. a. Distribution of ESP homologs in Asgard archaea grouped by function. Shaded rectangles above protein names indicate ESPs newly identified as part of this study. Predicted homologs are depicted by colored circles: fully filled circles indicate 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 previously¹⁰. Figure generated with Biorender.com.

lowest OGT among Heimdallarchaeia. c. Based on the presence/absence of the thermophily-
diagnostic enzyme reverse gyrase and the following metabolic signatures in each of the ancestors,
we predict that the last Asgard common ancestor probably transitioned from a hyperthermophilic
fermentative lifestyle to a mesophilic mixotroph lifestyle. The LAsCA likely encoded
gluconeogenic pathways *via* the reverse EMP gluconeogenic pathway and *via* FBP
aldolase/phosphatase (FBP A/P). The major energy-conserving step in the early Asgard ancestors
could have been the ATP synthesis by fermentation of small organic molecules (i.e., acetate,
formate, formaldehyde). The reverse ribulose monophosphate pathway (rRuMP) was a key
pathway in the LAsCA for the generation of reducing power. The Wood-Ljungdahl pathway
(WLP) appeared only to be present in the LAsCA and was lost in the more recent ancestors (of
Heimdallarchaeia and Hodarchaeales) indicated here. The tricarboxylic acid (TCA) cycle is
predicted to be complete in all five investigated Asgard ancestors. The inferred presence of the
electron transport chain (ETC) components is shown for selected ancestors of major Asgard
archaea groups, with the Hodarchaeales common ancestor encoding the most complete set of ETC
subunits, and likely using nitrate as a terminal electron acceptor. Therefore, membrane-associated
ATP biosynthesis coupled to the oxidation of NADH and succinate and reduction of nitrate to
nitrite within the respiratory chain could have been present in the LAECA. Abbreviations: Q:
quinone; c: cupredoxin; FBP A/P: Fructose 1,6-bisphosphate aldolase/phosphatase; EMP:
Embden-Meyerhof-Parnas; OPPP: Oxidative pentose phosphate pathway; rRuMP: Reversed
ribulose monophosphate pathway; RHP: reductive hexulose-phosphate; RuBisCO: Ribulose-1,5-
bisphosphate carboxylase/oxygenase; PRK: phosphoribulokinase; AMP: adenosine
monophosphate salvage pathway. Details of copy numbers of key enzymes involved in central
carbon metabolism are found in Supplementary Table 4.

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 -80°C 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 v2.3.5.1⁶⁵. The coverage of a given scaffold was calculated using the command of “jgi_summarize_bam_contig_depths” in MetaBAT v2.12.1⁶⁶. 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

considered. The clustering of scaffolds from the bins and the unbinned scaffolds was visualized using ESOM with a minimum window length of 2.5 kbp and max window length of 5 kbp, as previously described⁶⁷. Misplaced scaffolds were removed from bins and unbinned scaffolds whose segments were placed within the bin areas of ESOMs were added to the corresponding bins. Scaffolds with a minimum length of 1 kbp were uploaded to ggKbase (<http://ggkbase.berkeley.edu/>). The ESOM-curated bins were further evaluated based on consistency of GC content, coverage and taxonomic information, and scaffolds identified with abnormal information were removed. The ggKbase genome bins were curated individually to fix local assembly errors using ra2.py⁶⁸.

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

Hydrothermal vent deposits were collected from the ABE (ABE 1, 176° 15.48'W, 21° 26.68'S, 2142 m; ABE 3, 176° 15.59'W, 21° 26.95'S, 2131 m) and Mariner (176° 36.07'W, 22° 10.81'S, 1914 m) vent fields along the Eastern Lau Spreading Center in April/May of 2015 during the RR1507 Expedition on the RV Roger Revelle. Sample collection and processing were done as previously described⁶⁹. DNA was extracted from homogenized rock slurries using the DNeasy PowerSoil kit (Qiagen) as per the manufacturer's instructions. Samples were prepared for sequencing on the Illumina HiSeq 3000 using Nextera DNA Library Prep kits (Illumina), and metagenomes (2x150 bp) were sequenced at the Oregon State University Center for Genome Research and Computing. Trimmomatic⁷⁰ v.0.36 was used to trim low-quality regions and adapter sequences from raw reads (parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10, LEADING:20, SLIDINGWINDOW:4:20, MINLEN:50). Clean paired reads were then interleaved using the khmer software package⁷¹. Interleaved and unpaired reads were assembled with MEGAHIT v.1.1.1-2-g02102e1 (--k-min 31, --k-max 151, --k-step 20, --min-contig-len 1000)

^{72,73}. Trimmed reads were mapped back to the contigs to determine read coverage using Bowtie 2 v.2.2.9^{65,74} and SAMtools v.1.3.1⁷⁵. Binning was performed with MetaBAT v.0.32.4⁶⁶ using tetranucleotide frequency and read coverage. Bin completion and contamination were estimated with CheckM v.1.0.7⁷⁶.

c. Lokiarchaeote ABR01, ABR02, ABR03, ABR04, ABR05, ABR06, ABR08, ABR11, ABR13, ABR15, Thorarchaeote ABR09, ABR10 and Heimdallarchaeote ABR14, ABR16 MAGs were obtained as previously described³¹.

d. Archaeon WORA1, WORB2, Heimdallarchaeote WORE3, Lokiarchaeote WORB4, WORC5 and Thorarchaeote WORH6

Sampling, DNA extraction, sequencing library preparation and sequencing methods were previously described⁷⁷. Published assemblies and raw reads for the samples WOR-1-36_30 (SAMN06268458; Gp0056175), WOR-1-52-54 (SAMN06268416; Gp0059784), WOR-3-24_28 (SAMN06268417; Gp0059785) were downloaded from JGI. Short reads were trimmed using Trimmomatic⁷⁰ v0.33 (PE ILLUMINACLIP:2:30:10 SLIDINGWINDOW:4:15 MILEN:100). Contigs shorter than 1000 bp were excluded from the assembly using SeqTK v1.0r75 (<https://github.com/lh3/seqtk>). Each assembly was binned using CONCOCT v0.4.1⁷⁸ and coverage information from the three datasets, and Asgard bins were subsequently identified based on phylogenies of concatenated ribosomal proteins¹⁰. Identified Asgard MAGs were used together with publicly available Asgard genomes to recruit trimmed-reads originated from Asgard genomes using CLARK v1.2.3 with the -m 0 option⁷⁹. For each dataset, recruited Asgard reads were independently assembled using SPAdes⁸⁰ and IDBA-UD⁸¹ and further binned using CONCOCT, using a minimum contig length of 1000 bp. Bins with higher completeness and lower contamination values as predicted by miComplete v1.00⁸² were selected and manually curated

using mmgenome v0.7.1^{83,84} using the coverage information, paired-reads linkage, composition and marker genes information. The samples and assembly method used for each final MAG were: Archaeon WORA1 (WOR-1-52-54; spades), Archaeon WORB2 (WOR-1-52-54; IDBA-UD), Heimdallarchaeote WORE3 (WOR-3-24_28; spades), Lokiarchaeote WORB4 and WORC5 (WOR-1-36_30; IDBA-UD), and Thorarchaeote WORH6 (WOR-1-36_30; spades).

e. Jordarchaeote RPD1, RPF2 and Odinararchaeote RPA3

Information about the location of the hot spring sediments from Radiata Pool, sampling and DNA extraction procedures was previously reported¹⁰. Short paired-end Illumina reads were generated and preprocessed using Scythe (<https://github.com/vsbuffalo/scythe>) and Sickle (<https://github.com/najoshi/sickle>) to remove adapters and low-quality reads. Reads were subsequently assembled of IDBA-UD 1.1.3 (--maxk 124). The Jordarchaeote RPF2 MAG was generated by binning contigs according to their tetranucleotide frequencies using esomWrapper.pl (<https://github.com/tetramerFreqs/Binning>) with a minimum contig length 5000 bp and a window size of 10 Kbp. ESOM maps were manually delineated using the Databionic ESOM viewer (<http://databionic-esom.sourceforge.net/>). Jordarchaeote RPD1 and Odinararchaeote RPA3 were binned following the methodology described in above (section d), but re-assembling the recruited reads only with IDBA-UD (--maxk 124)⁸¹.

f. Jordarchaeote GBS01, GBS02, GBS03, GBS04, GBS05, GBS06, GBS07, Heimdallarchaeote GBS08, GBS09, GBS10, GBS11, Lokiarchaeote GBS14, Njordarchaeote GBS15, GBS16, GBS17, GBS18, GBS19, GBS20, GBS21, GBS22, GBS23, GBS24, GBS25, GBS26, TNS08 and Thorarchaeote GBS28, GBS29, GBS33, GBS34 MAGs were obtained as described in ⁸⁵.

g. Heimdallarchaeote B3_JM_08 MAG was obtained as described in ⁸⁶.

h. Thorarchaeote OWC_bin2, OWC_bin3 and OWC_bin4 MAGs were obtained as described in ³¹.

i. Heimdallarchaeote GBS11

Samples were made available by the Gulf Coast Repository (GCR) and were collected on the Ocean drilling Program (ODP) Leg 204 at site 1244 (44°35.17N, 125°7.19W) on July 14th, 2002 (hole C and core 2). The ODP site is found at a water depth of 890 m on the eastern side of the South Hydrate Ridge on the Cascadia Margin. This site has been well characterized physically and geochemically⁸⁷. Furthermore, the microbial community structure has been surveyed using 16S rRNA gene sequencing^{88,89}. Two sediment samples, designated DCO-2-5 (sample ID 1489929) and DCO-2-7 (sample ID 1489924), were collected at a sediment depth of 12.40 and 14.96 m below the seafloor, respectively, and stored at -80°C at GCR. A total amount of 10 g of each of the two sediment samples was used to extract DNA using the MoBio DNA PowerSoil Total kit. A total amount of 100 ng DNA was used to prepare sequencing libraries that were 150 bp paired-end sequenced at the Marine Biological Laboratory (Woods Hole, MA, USA) on an Illumina MiSeq sequencer. Adaptors and DNA spike-ins were removed from the forward and reverse reads using cutadapt v1.12⁹⁰. Afterwards, reads were interleaved using [interleave_fasta.py](https://github.com/jorvis/biocode/blob/master/fasta/interleave_fasta.py) (https://github.com/jorvis/biocode/blob/master/fasta/interleave_fasta.py), and further trimmed using Sickle with default settings (Fass JN) (<https://github.com/najoshi/sickle>). Metagenomic reads from both samples were co-assembled using IDBA-UD using the following parameters: --pre_correction, -mink 75, -maxk 105, --step 10, --seed_kmer 55⁸¹. Metagenomic binning was performed on scaffolds with a length >3,000 bp using ESOM, including a total of 4,939 scaffolds with a length of 30,693,002 bp^{67,81}. CheckM v1.0.5 was employed to evaluate the accuracy of the binning approach by determining the percentage of completeness and contamination⁷⁶.

j. Heimdallarchaeote GBS09

MAG was obtained as previously described⁹¹.

Exploration of phylogenetic diversity in Asgard assemblies and MAGs

To assess the presence of potential Asgard-related lineages in our assemblies, we reconstructed a phylogeny of ribosomal proteins encoded in a conserved 15-ribosomal protein (RP15) gene cluster¹⁶. As ingroup, we used all MAGs presented in this study, plus all genomes classified as Asgard archaea in NCBI as of June 25th 2021, plus those classified as “archaeon” corresponding to Hermodarchaeia (GCA_016550385.1, GCA_016550395.1, GCA_016550405.1, GCA_016550415.1, GCA_016550425.1, GCA_016550485.1, GCA_016550495.1, GCA_016550505.1). and all Asgard archaeal MAGs released by Sun et al.²¹. To obtain an adequate outgroup dataset, we downloaded all archaeal genomes from the Genome Taxonomy Database⁹², data revision 89, and selected one genome sequence per species-level cluster as defined in https://data.gtdb.ecogenomic.org/releases/release89/89.0/sp_clusters_r89.tsv. We then selected a set of 216 genomes classified as Bathyarchaeia, Nitrososphaeria and Thermoprotei, and used them as outgroup. Genes were detected and individually aligned and trimmed as previously described¹⁰. Ribosomal protein sequences were selected if they were encoded in a contig containing at least five of the 15 ribosomal protein genes. ModelFinder⁹³ was run as implemented in IQ-TREE v. 2.0-rc2 to identify the best model among all combinations of the LG, WAG, JTT and Q.pfam models, as well as their corresponding mixture models by adding +C20, +C40 and +C60, and the additional mixture models LG4M, LG4X, UL2 and UL3, with rate heterogeneity (none, +R4 and +G4) and frequency parameters (none, +F). A PMSF approximation⁹⁴ of the chosen model (WAG+C60+R4+F) was then used for a final reconstruction using 100 non-

parametric bootstrap pseudoreplicates for branch statistical support. The obtained tree revealed a broad genomic diversity of Asgard lineages (Figure 1).

Environmental distribution of Asgard archaea

16S rRNA gene sequences were predicted with Barrnap v 0.9 (<https://github.com/tseemann/barrnap>) with the option “--kingdom arc”. Since none of the two Helarchaeales bins contained 16S rRNA gene sequences, helarchaeal 16S rRNA gene sequences identified by Seitz et al.⁵⁰ were used as representatives of this phylum. These sequences were submitted to the IMNGS platform as queries for Paraller similarity searches against all available NCBI sequence read archive (SRA) samples with a 95% similarity threshold and a minimum alignment size of 200 bp⁹⁵. Available metadata for detected SRA samples were then manually assessed to link environmental context descriptions for individual SRA samples to broader environment categories. The sequence abundance output file generated by IMNGS was then analysed using R and the package tidyverse to calculate the number of SRA samples belonging to each environment per phylum⁹⁶.

Gene prediction

Gene prediction was performed using Prokka⁹⁷ v1.12 (prokka --kingdom Archaea --norrna --notrna). rRNA genes and tRNA genes were predicted with Barrnap (<https://github.com/tseemann/barrnap>) and tRNAscan-SE^{96,98}, respectively.

944

945 **Optimal growth temperature prediction**

946 Optimal growth temperatures were predicted for the genomes presented here based on genomic
947 and proteomic features⁵⁶ (see Supplementary Information). Since ribosomal RNAs nucleotide
948 composition are used in this method, only genomes with predicted rRNAs were analyzed.

949

950 **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 diamond⁹⁹ blastp (--more-sensitive --evaluate 0.0001 --max-target-seqs 0 --outfmt
954 6). The file generated was used to cluster protein sequences into homologous families using
955 SiLiX¹⁰⁰ v.1.2.10, followed by Hifix¹⁰¹ 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

960 **Functional annotation of homologous protein families**

961 Protein families, excluding singletons, were aligned using mafft-linsi¹⁰² v7.402 and converted into
962 HHsearch format (.hhm) profiles using HHblits¹⁰³ v3.0.3. Profile-profile searches were
963 subsequently performed against a database containing profiles from EggNOG 4.5¹⁰⁴, arCOGs¹⁰⁵
964 and PFAM databases¹⁰⁶ that had been previously converted to the hhm format using HHblits¹⁰³
965 v3.0.3.

966

967 **Automatic functional annotation of individual proteins**

968 Individual proteins were annotated using the HMMscan tool of the HMMer suite against PFAM
 969 v32¹⁰⁶, Interproscan¹⁰⁷ 5.25-64.0, EggNOG mapper v0.12.7 against the NOG database v4.5¹⁰⁴,
 970 diamond aligner v0.9.9.110⁹⁹ against the *nr* database, arCOG¹⁰⁵, and GhostKoala annotation
 971 server¹⁰⁸. Putative hydrogenases were further classified using HydDB¹⁰⁹.

972

973 **Detailed analysis of ESPs**

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

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

strict cutoff that the diagnostic HMM profile had the best score among all profiles detected for a protein. An exception was made for the ESCRT domain Vps28, Steadiness box, UEV, Vps25, NZF, GLUE and Vps22 domains which are usually found in combination with other protein domains and thus do not necessarily represent the best scoring domain in a protein even if they represent true homologs.

Phylogenetic analyses of concatenated proteins for species tree inference

Two sets of phylogenetic markers were used to infer the species tree. The first one (RP56) is based on a previously published dataset of 56 ribosomal proteins used to place the first assembled Asgard genomes¹⁰. The second one (NM57, for ‘new markers’) corresponds to 57 proteins extracted from a set of 200 markers previously identified as core archaeal proteins that can be used to robustly infer the tree of archaea¹¹². These 57 markers were selected because they were found in at least a third of representatives of each of the 11 Asgard clades, as well as in 10 out of 14 eukaryotes, and were inherited from archaea in eukaryotes.

We initially assembled an RP56 dataset for a phylogenetically diverse set of 222 archaeal and 14 eukaryotic taxa. These included all 11 Asgard archaea MAGs and genomes available at the NCBI as of May 12, 2017, as well as the 53 most diverse novel MAGs from this work (out of 63). We gathered orthologs of these genes from all proteomes by using sequences from the previously published alignment^{10,112} as queries for BLASTp. For each marker, the best BLAST hit from each proteome was added to the dataset. For the first iteration, each dataset was aligned using mafft-linsi¹¹³ and ambiguously aligned positions were trimmed using BMGE (-m BLOSUM30)¹¹⁴. All 56 trimmed RP alignments were concatenated into an RP56-A64 supermatrix (236 taxa including

64 Asgard archaea, 6332 amino acid positions). Once this taxon set was gathered, we identified homologs of the NM57 gene set as described above, thus generating supermatrix NM57-A64 (236 taxa, 14,847 amino acid positions).

We carried out a large number of phylogenomic analyses on variations of these two RP56-A64 and NM57-A64 datasets with different phylogenetic algorithms. Notably, preparing these datasets must be done with great care and is therefore time-consuming, and subsequent phylogenomic analyses generally require an enormous amount of computational running time. However, the rapid expansion of available Asgard archaeal MAGs, notably by Liu and colleagues as of April 2021¹⁵, urged us to update and re-run many of the computationally demanding analyses. As some of the work that was based on a more restrained taxon sampling is still deemed valuable, such as some of the Bayesian phylogenomic analyses and ancestral genome content reconstructions, we retained these in the present study.

An updated Asgard archaeal genomic sequence dataset was constructed by including all 230 Asgard archaeal MAGs and genomes available at the NCBI database as of May 12, 2021, as well as 63 novel MAGs described in the present work. All 56 trimmed RP alignments were concatenated into an RP56-A293 supermatrix (465 taxa including 293 Asgard archaea, 7112 amino acid positions), which was used to infer a preliminary phylogeny with FastTree v2¹¹⁵ (Supplementary Figure 16). Given the high computational demands of the subsequent analyses, we then used this phylogeny to select a subsample of Asgard archaea representatives. For this, we first removed the most incomplete MAGs encoding fewer than 19 ribosomal proteins (i.e., 1/3 of the markers) in the matrix. We also used the preliminary phylogeny to sub-select among closely related taxa: among taxa that were separated by branch lengths of <0.1, we only kept one representative. This led to a selection of 331 genomes, including 175 Asgard archaea, 41 DPANN,

43 Euryarchaeota, and 72 TACK representatives (RP56-A175 dataset). Out of these 175 Asgard archaea, 41 correspond to MAGs newly reported here. Once this taxon set was gathered, we identified homologs of the NM57 gene set as described above, thus generating supermatrix NM57-A175. All datasets and their composition are summarized in Supplementary Table 2.

To test for potential phylogenetic reconstruction artefacts, our datasets were subjected to several treatments. Supermatrices were recoded into four categories, using the SR4 scheme²⁷. The corresponding phylogenies were reconstructed with IQ-TREE (using a user-defined previously described model referred to as ‘C60SR4’, based on the implemented ‘LG+C60’ model and modified to analyze the recoded data¹⁰) and Phylobayes (under the CAT+GTR model)¹⁰. We also used the estimated site rate output generated by IQ-TREE (-wsr) to classify sites into 10 categories, from the fastest to the slowest evolving, and we removed them in a stepwise fashion, removing from 10% to 90% of the data. Finally, we combined both approaches by applying SR4 recoding to the alignments obtained after each fast-site removal step. All phylogenetic analyses performed are summarized in Supplementary Table 2. See Supplementary Information for details and discussion.

Analyses of individual proteins

For individual proteins of interest, we gathered homologs using various approaches, depending on the level of conservation across taxa. In order to detect putative Asgard homologs of eukaryotic proteins, we used a combination of tools including BLASTp¹¹⁶ and the HMMer toolkit (<http://hmmer.org/>) if HMM profiles were available, and queried a local database containing our 240 archaeal representatives (including all Asgard predicted proteomes). We then investigated the Asgard candidates by 1) using them as seed for BLASTp searches against nr; 2) by 3D modelling

using Phyre2 and Swissmodel when sequence similarity was low; 3) by annotating them using Interproscan 5.25-64.0¹⁰⁷, EggNOG mapper v0.12.7¹¹⁷, against the NOG database¹¹⁷, and GhostKoala annotation server¹⁰⁸; 4) by annotating the archaeal orthologous cluster they belonged to using profile-profile annotation as described above. Eukaryotic homologs were gathered from the UniRef50 database¹¹⁸. Depending on the divergence between homologs, they were aligned using mafft-linsi and trimmed using TrimAl¹¹⁹ (--automated1) or BMGE¹¹⁴, or, in cases where we investigated a specific functional domain, we used the hmmlalign tool from the HMMer package with the --trim flag to only keep and align the region corresponding to this domain. When divergence levels allowed, phylogenetic analyses were performed using IQ-TREE with model testing including the C-series mixture models (-mset option)¹²⁰. Statistical support was evaluated using 1000 ultrafast bootstrap replicates (for IQ-TREE)¹¹⁹.

Ancestral reconstruction

For the ancestral reconstruction analyses, only a subset of 181 taxa were included (64 Asgard, 74 TACK and 43 Euryarchaeota, see Supplementary Table 2 for details). Protein families with more than three members were aligned and trimmed using mafft-linsi v7.402¹¹³ and trimAl v1.4.rev15 with the --gappyout option¹¹⁹. Tree distributions for individual protein families were estimated using IQ-TREE v1.6.5 (-bb 1000 -bnni -m TESTNEW -mset LG -madd LG+C10,LG+C20 -seed 12345 -wbtl -keep-ident)¹²². The species phylogeny together with the gene tree distributions were subsequently used to compute 100 gene-tree species tree reconciliations using ALEobserve v0.4 and ALEml_undated^{52,53}, including the fraction_missing option that accounts for incomplete genomes. The genome copy number was corrected to account for the extinction probability per cluster (github.com/maxemil/ALE/commit/136b78e). The missing fraction of the genome was

calculated as 1 minus the completeness values (in fraction) as estimated by CheckM v1.0.5 for each of the 181 taxa⁷⁶. Protein families containing only one protein (singletons) were considered as originations at the corresponding leaf. The ancestral reconstruction of 5 protein families that included more than 2000 proteins raised errors and could not be computed. The minimum threshold of the raw reconciliation frequencies for an event to be considered was set to 0.3 as commonly done^{123–126} and recommended by the authors of ALE (Gergely Szöölősi, personal communication).

Ancestral metabolic inferences

Metabolic reconstruction of the Asgard ancestors was based on the inference, annotation and copy number of genes in ancestral nodes. The presence of a given gene was scored if its copy number in the ancestral nodes was above 0.3. A protein family was scored as “maybe present” if the inferred copy number was between 0.1 and 0.3. The protein annotation of each of the clusters containing the ancestral nodes was manually verified for each of the enzymatic steps involved in the pathways detailed in Supplementary Table 4.

Data availability

The MAGs reported in this study have been deposited at DDBJ/EMBL/GenBank. BioProject IDs, BioSample IDs and GenBank assembly accession numbers are available in Supplementary Table 1 and will be released upon publication of the manuscript. All raw data underlying phylogenomic 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

1102 **Methods references**

1103 **Acknowledgements**

1104 We thank Stephan Köstlbacher for intellectual input. We thank the Uppsala Multidisciplinary
1105 Center for Advanced Computational Science (UPPMAX) at Uppsala University and the Swedish
1106 National Infrastructure for Computing (SNIC) at the PDC Center for High-Performance
1107 Computing for providing computational resources. We thank the Japan Agency for Marine-Earth
1108 Science & Technology (JAMSTEC) for taking sediment samples from the Taketomi shallow
1109 submarine hydrothermal system, the crew of the RV Roger Revelle for assisting with the sampling
1110 of the ABE and Mariner vent fields along the Eastern Lau Spreading Center during the RR1507
1111 Expedition. The Ngāti Tahu - Ngāti Whaoa Runanga Trust is acknowledged as *mana whenua* of
1112 Radiata Pool and associated samples, and we thank them for their assistance in access and sampling
1113 of the Ngatamariki geothermal features. Sampling in the Eastern Lau Spreading Center and
1114 Guaymas Basin (Gulf of California) was supported by the US-National Science Foundation (NSF-
1115 OCE-1235432 to A.-L. R. and NSF-OCE-0647633 to A.T.). A subset of Guaymas sediments were
1116 sequenced by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science
1117 User Facility under Contract No. DE-AC02-05CH11231 granted to ND. We thank the captain and
1118 crew of RV Aurora for assistance during sampling at Aarhus Bay. Sampling at Aarhus Bay was
1119 supported by the VILLUM Experiment project “FISHing for the ancestors of the eukaryotic cell”
1120 (grant number 17621 to A.S. and K.U.K.). This work was supported by grants of the European

Research Council (ERC Starting and Consolidator grants 310039 and 817834, respectively), the Swedish Research Council (VR grant 2015-04959), the Dutch Research Council (NWO-VICI grant VI.C.192.016), Marie Skłodowska-Curie ITN project SINGEK (H2020-MSCA-ITN-2015-675752) and the Wellcome Trust foundation (Collaborative award 203276/K/16/Z) awarded to T.J.G.E.. L.E. was supported by a Marie Skłodowska-Curie IEF (grant 704263), and by funding from the European Research Council (ERC Starting grant 803151). T.N. was supported by JSPS KAKENHI JP19H05684 within JP19H05679. W.-J.L. was supported by the National Natural Science Foundation of China (grant no. 91951205 and 92251302). D.T. was supported by the Swedish Research Council (International Postdoc grant 2018-06609). C.W.S. was supported by a Science for Life Laboratory postdoctoral fellowship (awarded to T.J.G.E. and C.W.S.) and funding from the Swedish research council (Vetenskapsrådet Starting grant 2020-05071 to C.W.S.). J.L. was supported by the Wenner-Gren Foundation (fellowship 2016-0072). J.H.S. was supported by a Marie Skłodowska-Curie IIF grant (331291). This work was also supported by the Moore-Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation 73592LPI to T.J.G.E. and B.J.B. (<https://doi.org/10.46714/735925LPI>) and Simons Foundation 812811 to L.E. (<https://doi.org/10.46714/735923LPI>), and NSF Division of Biological Science SBS Biodiversity: Discovery and Analysis program (1753661) to B.J.B.

Author contributions

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., A.-L.R., M.B.S., and A.P.T. collected and provided environmental samples. E.F.C., F.H., J.H.S., 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

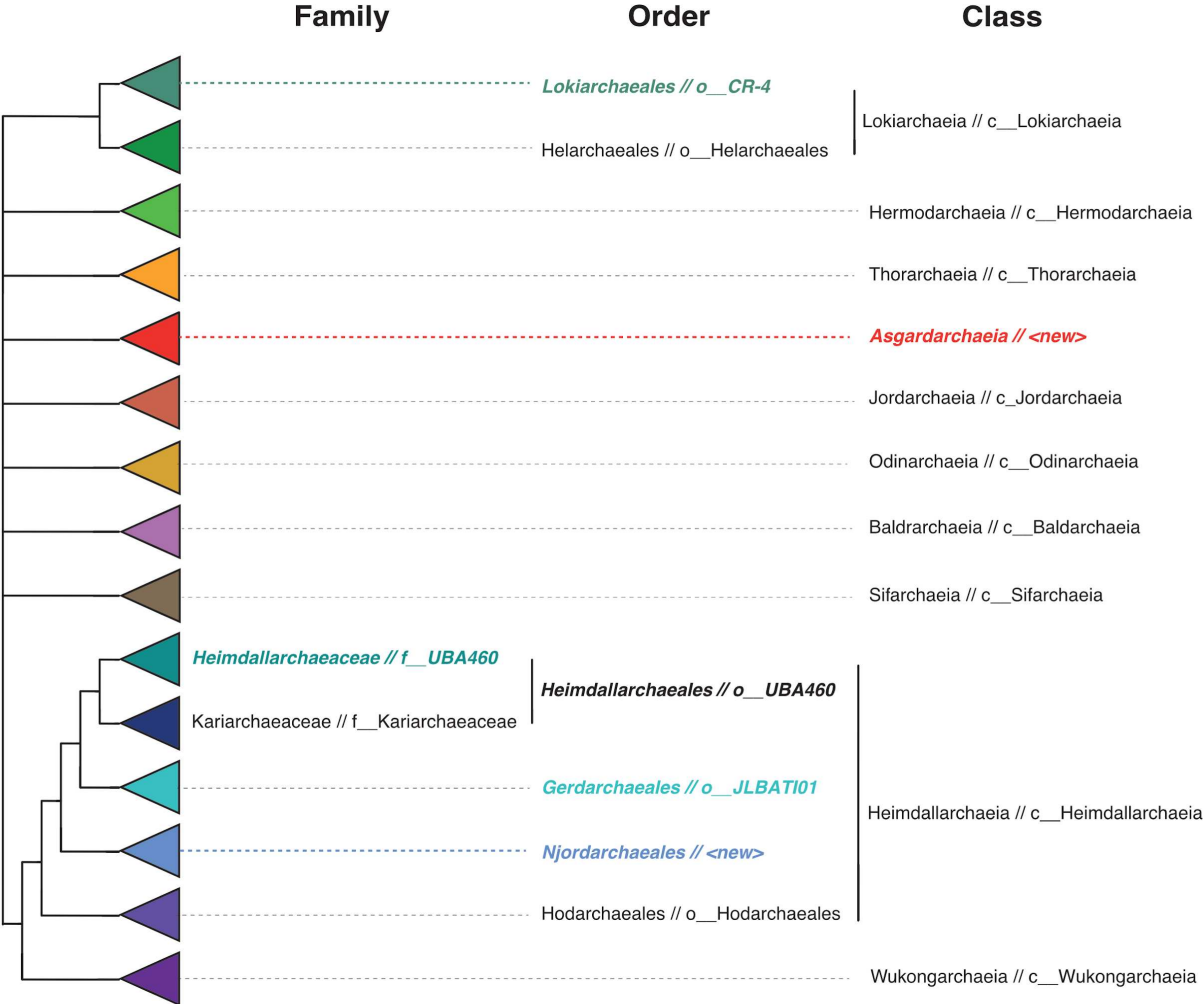
1151 **Competing interest declaration**

1152 The authors declare no competing financial interests.

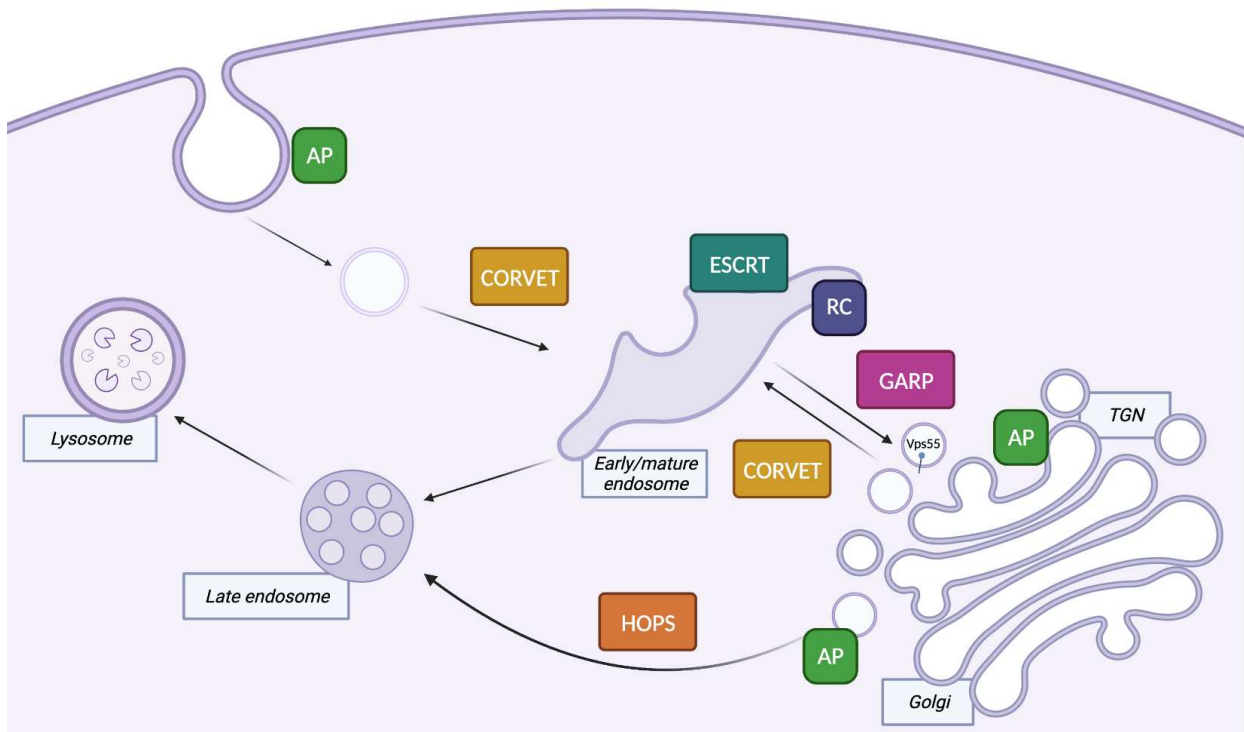
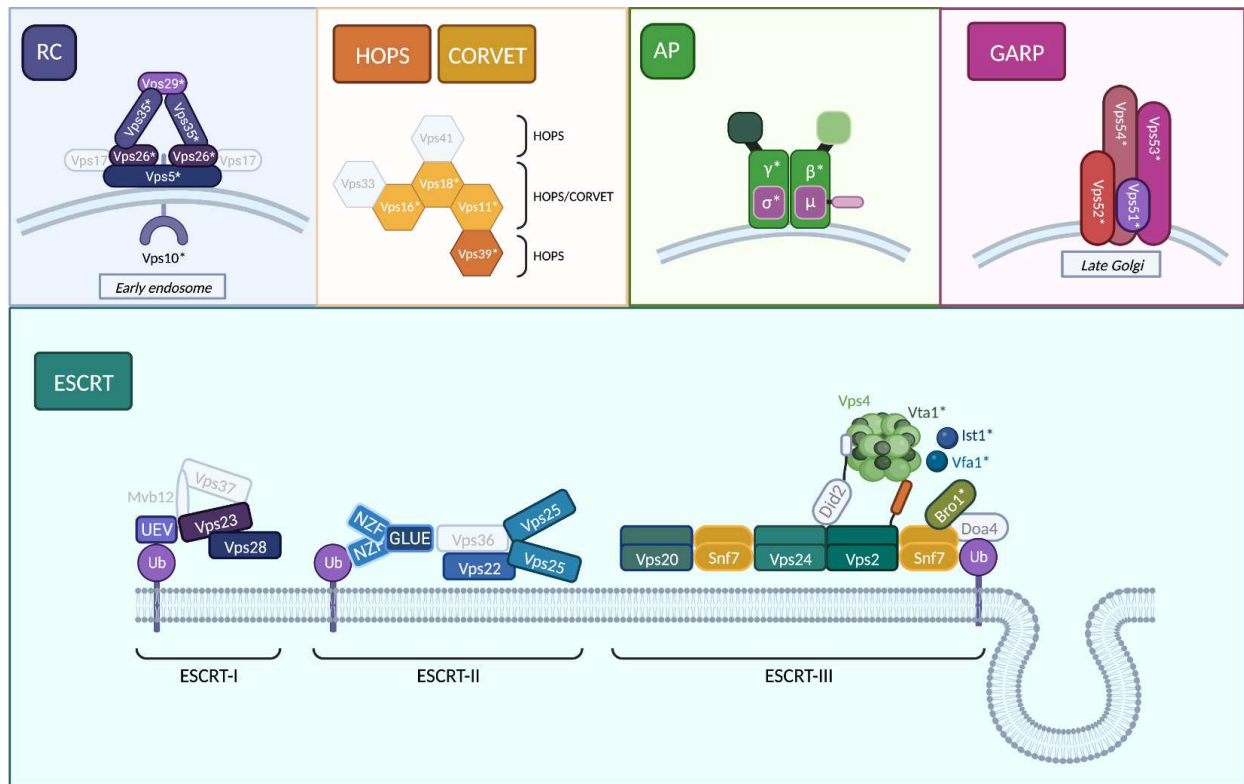
1153

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

Extended Data Figure 2. Identification of previously undetected vesicular trafficking ESPs in Asgard archaea. Schematic representation of a eukaryotic cell in which ESPs involved in membrane trafficking and endosomal sorting that have been identified in Asgard archaea are highlighted. Colored subunits have been detected in some Asgard archaea while grey ones seem to be absent from all current representatives. Only major protein complexes are depicted. Additional components can be found in Figure 2. From left to right, top to bottom: RC, Retromer complex. Retromer is a coat-like complex associated with endosome-to-Golgi retrograde traffic³⁶. It is formed by Vacuolar protein sorting-associated protein 35, Vps5, Vps17, Vps26 and Vps29¹²⁷. During cargo recycling, retromer is recruited to the endosomal membrane via the Vps5-Vps17 dimer. Cargo recognition is thought to be mediated primarily through Vps26 and possibly by Vps35. Finally, the BAR domains of Vps5-Vps17 deform the endosomal membrane to form cargo-containing recycling vesicles. Their distribution is sparse, but we have detected Asgard archaeal homologs of all subunits except for Vps17. Interestingly, the Thorarchaeota Vps5-BAR domain is often fused to Vps28, a subunit of the ESCRT machinery complex I, suggesting a functional link between BAR domain proteins and the thorarchaeal ESCRT complex. The best-characterized retromer cargo is Vps10. This transmembrane protein receptor is known in yeast and mammal cells to be involved in the sorting and transport of lipoproteins between the Golgi and the endosome. The Vps10 receptor releases its cargo to the endosome and is recycled back to the Golgi via the retromer complex¹²⁸. CORVET: Class C core vacuole/endosome tethering complex; HOPS: Homotypic fusion and protein sorting complex. Endosomal fusion and autophagy depend on the CORVET and HOPS hexameric complexes³⁹; they share the core subunits Vps11, Vps16, Vps18, and Vps33⁴⁰. In addition, HOPS is composed of Vps41 and Vps39⁴¹. Vps39, found associated to late endosomes and lysosomes, promotes endosomes/lysosomes clustering and their

fusion with autophagosomes¹²⁹. AP, Adaptor Proteins. Asgard archaea genomes from diverse phyla encode key functional domains of the AP complexes. The eukaryotic AP tetraheteromeric structure is depicted, each color corresponding to a PFAM functional domain (Medium green: Adaptin, N terminal region; Dark green: Alpha adaptin, C-terminal domain; Light green: Beta2-adaptin appendage, C-terminal sub-domain; Dark pink/clear outline: Clathrin adaptor complex small chain; Light pink/dark outline: C-ter domain of the mu subunit); all five domains were detected in Asgard archaea, although not fused to each other. GARP: Golgi-associated retrograde protein complex. The GARP complex is a multisubunit tethering complex located at the trans-Golgi network where it functions to tether retrograde transport vesicles derived from endosomes^{37,38}. GARP comprises four subunits, VPS51, VPS52, VPS53, and VPS54. ESCRT: Endosomal Sorting Complex Required for Transport system. This complex machinery performs a topologically unique membrane bending and scission reaction away from the cytoplasm. While numerous components of the ESCRT-I, II and III systems have been previously detected in Asgard archaea^{9,10,42}, we here report Asgard homologs for several ESCRT-III regulators Vfa1, Vta1, Ist1, and Bro1. The bottom panel shows where these complexes mainly act in eukaryotic cells. Ub: Ubiquitin; Vps: vacuolar protein sorting. Subunit names in grey indicate that no homologs were detected in Asgard archaea. Domains newly identified as part of this study are indicated with an asterisk. Created with BioRender.com.

1209 **Supplementary information**

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.