# Expanding diversity of Asgard archaea and the elusive ancestry of eukaryotes

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

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

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

3 State Key Laboratory Breeding Base of Marine Genetic Resources, Key Laboratory of Marine Genetic Resources, Fujian Key Laboratory of Marine Genetic Resources, Third Institute of Oceanography, State Oceanic Administration, Xiamen 361005, P. R. China

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

† These authors contributed equally to this work.

*Authors for correspondence: koonin@ncbi.nlm.nih.gov or limeng848@szu.edu.cn

Running title: Asgard archaea genomics

Keywords:

## Abstract

Comparative analysis of 162 (nearly) complete genomes of Asgard archaea, including 75 not reported previously, substantially expands the phylogenetic and metabolic diversity of the Asgard superphylum, with six additional phyla proposed. Phylogenetic analysis does not strongly support origin of eukaryotes from within Asgard, leaning instead towards a three-domain topology, with eukaryotes branching outside archaea. Comprehensive protein domain analysis in the 162 Asgard genomes results in a major expansion of the set of eukaryote signature proteins (ESPs). The Asgard ESPs show variable phyletic distributions and domain architectures, suggestive of dynamic evolution via horizontal gene transfer (HGT), gene loss, gene duplication and domain shuffling. The results appear best compatible with the origin of the conserved core of eukaryote genes from an unknown ancestral lineage deep within or outside the extant archaeal diversity. Such hypothetical ancestors would accumulate components of the mobile archaeal 'eukaryome' via extensive HGT, eventually, giving rise to eukaryote-like cells.

## Introduction

The Asgard archaea are a recently discovered archaeal superphylum that is rapidly expanding, thanks to metagenomic sequencing (*1*–*5*). The Asgard genomes encode a diverse repertoire of eukaryotic signature proteins (ESPs) that far exceeds the diversity of ESPs in other archaea. The Asgard ESPs are particularly enriched in proteins involved in membrane trafficking, vesicle formation and transport, cytoskeleton formation and the ubiquitin network, suggesting that these archaea possess a eukaryote-type cytoskeleton and an intracellular membrane system (*2*).

The discovery of the Asgard archaea rekindled the decades old but still unresolved fundamental debate on the evolutionary relationship between eukaryotes and archaea that has shaped around the '2-domain (2D) versus 3-domain (3D) tree of life' theme (*6*–*8*). The central question is whether the eukaryotic nuclear lineage evolved from a common ancestor shared with archaea, as in the 3D tree, or from within the archaea, as in the 2D tree. The discovery and phylogenomic analysis of Asgard archaea yielded strong evidence in support of the 2D tree, in which eukaryotes appeared to share common ancestry with one of the Asgard lineages, Heimdallarchaeota (*1*, *2*, *5*). However, the debate is not over as arguments have been made for the 3D topology, in particular, based on the phylogenetic analysis of RNA polymerases, some of the most highly conserved, universal proteins (*9*, *10*).

Molecular phylogenetic methods alone might be insufficient to resolve the ancient ancestral relationship between archaea and eukarya. To arrive at a compelling solution, supporting biological evidence is crucial (*11*), because, for example, the transition from archaeal, ether-linked membrane lipids to eukaryotic, ester-linked lipids that constitute eukaryotic (and bacterial) membranes (*12*) and the apparent lack of the phagocytosis capacity in Asgard archaea (*13*) are major problems for scenarios of the origin of eukaryotes from Asgard or any other archaeal lineage. Some biochemical evidence indicates that Asgard archaea possess an actin cytoskeleton regulated by accessory proteins, such as profilins and gelsolins, and the endosomal sorting complex required for transport machinery (ESCRT) that can be predicted to function similarly to the eukaryotic counterparts (*14*–*16*). Generally, however, the biology of Asgard archaea remains poorly characterized, in large part, because of their recalcitrance to growth in culture (*17*). To date, only one Asgard archaeon, *Candidatus* Prometheoarchaeum syntrophicum strain MK-D1, has been isolated and grown in culture (*17*). This organism has been reported to form extracellular protrusions that are involved in its interaction with syntrophic bacteria, but no visible organelle-like structure and, apparently, little intracellular complexity.

The only complete, closed genome of an Asgard archaeon also comes from *Candidatus* P. syntrophicum strain MK-D1 (*17*) whereas all other genome sequences were obtained by binning multiple metagenomics

contigs. Furthermore, these genome sequences represent but a small fraction of the Asgard diversity that has been revealed by 16S rRNA sequencing (*18*, *19*).An additional challenge to the study of the relationship between archaea and eukaryotes is that identification and analysis of the archaeal ESPs are non-trivial tasks due to the high sequence divergence of many if not most of these proteins. At present, the most efficient, realistic approach to the study of Asgard archaea and their eukaryotic connections involves obtaining high quality genome sequences and analyzing them using the most powerful and robust of the available computational methods.

Here we describe metagenomic mining of the expanding diversity of the superphylum Asgard, including the identification of six additional phylum-level lineages that thrive in a wide variety of ecosystems and are inferred to possess versatile metabolic capacities. We show that these uncultivated Asgard groups carry a broad repertoire of ESPs many of which have not been reported previously. Our in depth phylogenomic analysis of these genomes provides insights into the evolution of Asgard archaea but calls into question the origin of eukaryotes from within Asgard.

## Results

### Reconstruction of Asgard archaeal genomes from metagenomics data

We reconstructed 75 metagenome-assembled genomes (MAGs) from the Asgard superphylum that have not been reported previously. These MAGs were recovered from various water depths of the Yap trench, intertidal mangrove sediments of Mai Po Nature Reserve (Hong Kong, China) and Futian Mangrove Nature Reserve (Shenzhen, China), seagrass sediments of Swan Lake Nature Reserve (Rongcheng, China) and petroleum samples of Shengli oilfield (Shandong, China) (Supplementary Table 1, Supplementary Figure 1). For all analyses described here, these 75 MAGs were combined with 87 publicly available genomes, resulting in a set of 162 Asgard genomes. The 75 genomes reconstructed here were, on average, 82% complete and showed evidence of low contamination of about 3%, on average (Supplementary Figure 2).

### Classification of Asgard genes into clusters of orthologs

The previous analyses of Asgard genomes detected a large fraction of "dark matter" genes (*20*). For example, in the recently published complete genome of *Candidatus* Prometheoarchaeum syntrophicum, 45% of the proteins are annotated as "hypothetical". We made an effort to improve the annotation of Asgard genomes by investigating this dark matter in greater depth, and developing a dedicated platform

for Asgard comparative genomics. To this end, we constructed Asgard Clusters of Orthologous Genes (asCOGs) and used the most sensitive available methods of sequence analysis to annotate additional Asgard proteins, attempting, in particular, to expand the catalogue of Asgard homologs of ESPs (see Materials and Methods for details).

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

The gene commonality plot for the asCOGs shows an abrupt drop at the right end, which reflects a surprising deficit of nearly universal genes (Fig. 1). Such shape of the gene commonality curve appears anomalous compared to other major groups of archaea or bacteria with many sequenced genomes (*21*). For example, in the case of the TACK superphylum of archaea, for which the number of genomes available is similar to that for Asgard, with a comparable level of diversity, the commonality plot shows no drop at the right end, but instead, presents a clear uptick, which corresponds to the core of genes represented in (almost) all genomes (Fig. 1). Apparently, most of the Asgard genomes remain incomplete, such that conserved genes were missed randomly. Currently, there are only three gene families that are present in all Asgard MAGs, namely, a Zn-ribbon domain, a Threonyl-tRNA synthetase and an aminotransferase (additional data file 1).

We employed the asCOG profiles to annotate the remaining 86 Asgard MAGs, including those that were sequenced in the later stages of this work (Supplementary Table 1). On average, 89% of the proteins encoded in these genomes were covered by asCOGs (Supplementary Table 1). Thus, the asCOGs database appears to be an efficient tool for annotation and comparative genomic analysis of Asgard MAGs and complete genomes.

## Expanding the phylogenetic diversity of Asgard archaea

Phylogenetic analysis of the Asgard MAGs based on a concatenated alignment of 209 core asCOGs (see Methods and additional data file 2) placed many of the genomes reported here into the previously delineated major Asgard lineages (Fig. 2a, Supplementary Table 1, additional data file 2), namely, Thorarchaeota (n=20), Lokiarchaeota (n=18), Hermodarchaeota (n=9), Gerdarchaeota (n=3),

Helarchaeota (n=2), and Odinarchaeota (n=1). Additionally, we identified 6 previously unknown major Asgard lineages that appear to be strong candidates to become additional phyla (Fig. 2a and b, Supplementary Table 1; see also Taxonomic description of new taxa in the Supplementary Information and additional data file 1). A clade formed by As_085 and As_075 is a deeply branching sister group to the previously recognized Heimdallarchaeota(*2*). Furthermore, our phylogenetic analysis supported the further split of "Heimdallarchaeota" into 4 phylum-level lineages according to the branch length in the concatenated phylogeny (see Materials and Methods; see also Taxonomic description of new taxa in the Supplementary Information and additional data file 1). The putative phyla within the old Heimdallarchaeota included the previously defined Gerdarchaeota (*4*), and three additional phyla that could be represented by As_002 (LC2) , As_003 (LC3) and AB_125 (As_001), respectively. Another 3 previously undescribed lineages were related, respectively, to Hel-, Loki-, Odin- and Thorarchaeota. Specifically, As_181, As_178 and As_183 formed a clade that was deeply rooted at the Hel-Loki-Odin-Thor clade; As_129 and As_130 formed a sister group to Odinarchaeota; and a lineage represented by As_086 was a sister group to Thorarchaeota. These results were buttressed by the 16S rRNA gene phylogeny, comparisons of the mean amino acid identity and 16S rRNA sequence identity (Fig. 2b, Supplementary Figure 3, Supplementary Figure 4, Supplementary Table 2 and Supplementary Table 3).

We propose the name Wukongarchaeota after Wukong, a Chinese legendary figure who caused havoc in the heavenly palace, for the putative phylum represented by MAGs As_085 and As_075 (*Candidatus* Wukongarchaeum yapensis), and names of Asgard deities in the Norse mythology for the other 5 proposed phyla: (1) Hodarchaeota, after Hod, the god of darkness, for MAG As_027 (*Candidatus* Hodarchaeum mangrove); (2) Kariarchaeota, after Kari, the god of the North wind, for MAG As_030 (*Candidatus* Kariarchaeum pelagius); (3) Borrarchaeota after Borr, the creator god and father of Odin, for MAG As_133 (*Candidatus* Borrarchaeum yapensis); (4) Baldrarchaeota, after Baldr, the god of light and brother of Thor, for MAG As_130 (*Candidatus* Baldrarchaeum yapensis); (5) and Hermodarchaeota after Hermod, the messenger of the gods, son of Odin and brother of Baldr, for MAG As_086 (*Candidatus* Hermodarchaeum yapensis) (Fig. 2a). For details, see Taxonomic Description of new taxa in the **Supplementary Information.**

The gene content of Asgard MAGs agrees well with the phylogenetic structure of the group. The phyletic patterns of the asCOG form clusters that generally correspond to the clades identified by phylogenetic analysis (Fig. 3a), suggesting that gene gain and loss within Asgard archaea largely proceeded in a clock-like manner and/or that horizontal gene exchange preferentially occurred between genomes within the same clade.

**Phylogenomic analysis and positions of Asgard archaea and eukaryotes in the tree of life**

The outcome of phylogenetic reconstruction, especially, when deep branchings are involved, such as those that are relevant for the 3D vs 2D conundrum, depends on the phylogenetic methods employed, the selection of genes for phylogeny construction and, perhaps most dramatically, on the species sampling (*22–24*). In many cases, initially uncertain positions of lineages in a tree settle over time once more representatives of the groups in question and their relatives become available.

In our analysis of the universal phylogeny, we aimed to make the species set for phylogenetic reconstruction as broadly representative as possible, while keeping its size manageable, to allow the use of powerful phylogenetic methods. The tree was constructed from alignments of conserved proteins of 162 Asgard archaea, 286 other archaea, 98 bacteria and 72 eukaryotes (see Supplementary Material and Methods for details of the procedure including the selection of a representative species set and Supplementary Table 4). Members of 30 families of (nearly) universal proteins that appear to have evolved without much HGT and have been previously employed for the reconstruction of the tree of life (*25*) were used to generate a concatenated alignment of 7411 positions, after removing low information content positions (Supplementary Table 4). For the phylogenetic reconstruction, we used the IQ-tree program with several phylogenetic models (see Methods and Supplementary Table 5 for details). Surprisingly, the resulting trees had the 3D topology, with high support values for all key bifurcations (Fig. 2c, additional data file 2).

A full investigation of the effects of different factors, in particular, the marker gene selection, on the tree topology is beyond the scope of this work. Nonetheless, we addressed the possibility that the 3D topology resulted from the model used for the tree reconstruction and/or the species selection. To this end, we constructed 100 trees from the same alignment by randomly sampling 5 representatives of Asgard archaea, other archaea, bacteria and eukaryotes each. For these smaller sets of species, the best model identified by Williams et al. (LG+C60+G4+F) could be employed, resulting in 50 3D and 50 2D trees (Supplementary Table 5, additional data file 2). Because IQ-tree identified this model as over-specified for such a small alignment, we also tested a more restricted model (LG+C20+G4+F), obtaining 58 3D and 42 2D trees for the same set of 100 samples (Supplementary Table 5, additional data file 2).

The results of our phylogenetic analysis indicate that: 1) species sampling substantially affects the tree topology; 2) even the set of most highly conserved genes that appear to be minimally prone to HGT, yields conflicting signals for different species sets. Additional markers, less highly conserved and more prone to HGT, are unlikely to improve phylogenetic resolution and might cause systematic error. Notably, the topology of our complete phylogenetic tree (Fig. 2a) within the archaeal clade is mostly consistent

with the tree obtained in a preliminary analysis of a larger set of archaeal genomes and a larger marker gene set (*26*). Taking into account these observations and the fact that we used the largest set of Asgard archaea and other archaea compared to all previous phylogenetic analyses, the appearance of the 3D topology in our tree indicates that the origin of eukaryotes from within Asgard cannot be considered a settled issue. Various factors affecting the tree topology, including further increased species representation, particularly, of Asgard and the deep branches of the TACK superphylum, such as Bathyarchaeota and Korarchaeota, remain to be explored in order to definitively resolve the evolutionary relationship between archaea and eukaryotes.

**The core gene set of Asgard archaea**

We next analyzed the core set of conserved Asgard genes which we arbitrarily defined as all asCOGs that are present at least in one third of the MAGs in each of the 12 phylum-level lineages, with the mean representation across lineages >75%. Under these criteria, the Asgard core includes 378 asCOGs (Supplementary Table 6). As expected, most of these protein families, 293 (77%), are universal (present in bacteria, other archaea and eukaryotes), 62 (16%) are represented in other archaea and eukaryotes, but not in bacteria, 15 (4%) are found in other archaea and bacteria, but not in eukaryotes, 7 (2%) are archaea-specific, and only 1 (0.003%) is shared exclusively with eukaryotes (Supplementary Figure 5). Most of the core asCOGs show comparable levels of similarity to homologs from two or all three domains of life. The second largest fraction of the core asCOGs shows substantially greater sequence similarity (at least, 25% higher similarity score) to homologous proteins from archaea than to those from eukaryotes and/or bacteria (Supplementary Table 6). Compared with the 219 genes that comprise the pan-archaeal core (*27*), the Asgard core set lacks 12 genes, each of which, however, is present in some subset of the Asgard genomes. These include three genes of diphthamide biosynthesis and 2 ribosomal proteins, L40E and L37E. The intricate evolutionary history of gene encoding translation elongation factors and enzymes of diphthamide biosynthesis in Asgard has been analyzed previously (*28*). Also of note is the displacement of the typical archaeal glyceraldehyde-3-phosphate dehydrogenase (type II) by a bacterial one (type I) in most of the Asgard genomes (cog.001204, additional data file 1).

Functional distribution of the core asCOGs is shown in Fig. 3b (also see additional data file 1). For comparison, we also derived an extended gene core for the TACK superphylum, using similar criteria (at least 50% in each of the 6 lineages and 75% of the genomes overall, Fig 3b). For at least half of the Asgard core genes, across most functional classes, there were no orthologs in the TACK core. The most pronounced differences were found, as expected, in the category U (intracellular trafficking, secretion,

and vesicular transport). In Asgard archaea, this category includes 19 core genes compared with 7 genes in TACK; 13 of these genes are specific to the Asgard archaea and include components of ESCRT I and II, 3 distinct Roadblock/longin families, 2 distinct families of small GTPases, and a few other genes implicated in related processes (Supplementary Table 6).

We compared the protein annotation obtained using asCOGs with the available annotation of '*Candidatus* Prometheoarchaeum syntrophicum' and found that using asCOGs allowed at least a general functional prediction for 649 of the 1756 (37%) 'hypothetical proteins' in this organism, the only one in Asgard with a closed genome. We also identified 139 proteins, in addition to the 80 described originally, that can be considered Eukaryotic Signature Proteins, or ESPs (see next section).

## Eukaryotic features of Asgard archaea gleaned from genome analysis

The enrichment of Asgard proteomes with homologs of eukaryote signature proteins (ESPs), such as ESCRTs, components of protein sorting complexes including coat proteins, complete ubiquitin machinery, actins and actin-binding proteins gelsolins and profilins, might be the strongest argument in support of a direct evolutionary relationship between Asgard archaea and eukaryotes (*2*, *29*). However, the definition of ESPs is fuzzy because many of these proteins, in addition to their occurrence in Asgard, are either scattered among several other archaeal genomes, often, from diverse groups (*30*), or consist of promiscuous domains that are common in archaea, bacteria and eukaryotes, such as WD40 (after the conserved terminal amino acids of the repeat units, also known as beta-transducin repeats), LRR (Leucine-Rich Repeats), TPR (TetratricoPeptide Repeats), HEAT (Huntingtin-EF3-protein phosphatase 2A-TOR1) and other, largely, repetitive domains (*31*, *32*). Furthermore, the sequences of some ESPs have diverged to the extent that they become hardly detectable with standard computational methods. Our computational strategy for delineating an extensive yet robust ESP set is described under Materials and Methods. The ESP set we identified contained 505 asCOGs, including 238 that were not closely similar (E-value=$10^{-10}$, length coverage 75%) to those previously described by Zaremba-Niedzwiedzka et al. (*2*)(Supplementary Table 7). In a general agreement with previous observations, the majority of these ESPs, 329 of the 505, belonged to the 'Intracellular trafficking, secretion, and vesicular transport' (U) functional class, followed by 'Posttranslational modification, protein turnover, chaperones' (O), with 101 asCOGs (Supplementary Table 7). Among the asCOGs in the U class, 130 were Roadblock/LC7 superfamily proteins, including longins, sybindin and profilins, and 94 were small GTPases of several families, such as RagA-like, Arf-like and Rab-like ones, as discussed previously (*33*).

The phyletic patterns of ESP asCOGs in Asgard archaea are extremely patchy and largely lineage-specific (Fig. 4), indicating that most of the proteins in this set are not uniformly conserved throughout Asgard evolution, but rather, are prone to frequent HGT, gene losses and duplications. These evolutionary processes are correlated in prokaryotes, resulting in the overall picture of highly dynamic evolution (*34*). Even the most highly conserved ESP asCOG are missing in some Asgard lineages but show multiple duplications in others (Fig. 4 and Supplementary Table 7). Surprisingly, many gaps in the ESPs distribution were detected in the Heimdallarchaeota that include the suspected ancestors of eukaryotes.

Characteristically, many ESPs are multidomain proteins, with 37% assigned to more than one asCOG, compared to 17% among non-ESP proteins (Supplementary Table 7). Some multidomain ESPs in Asgard archaea have the same domain organizations as their homologs in eukaryotes, but these are a minority and typically contain only two domains. Examples include the fusion of two EAP30/Vps37 domains (*35*), and Vps23 and E2 domains (*35*) in ESCRT complexes, multiple Rag family GTPases, in which longin domain is fused to the GTPase domain, and several others. By contrast, most of the domain architectures of the multidomain ESP proteins were not detected in eukaryotes and often are found only in a narrow subset of Asgard archaea, suggesting extensive domain shuffling during Asgard evolution (Fig. 5a). For example, we identified many proteins containing a fusion of Vps28/Vps23 from ESCRT I complex (*35*) with C-terminal domains of several homologous subunits of adaptin and COPI coatomer complexes (*36*, *37*), and E3 UFM1-protein ligase 1, which is involved in the UFM1 ubiquitin pathway (*38*) (Fig. 5a). Generally, a protein with such a combination of domains can be predicted to be involved in ubiquitin-dependent membrane remodeling but, because its domain architecture is unique, the precise function cannot be inferred.

The majority of the ESP genes of Asgard archaea do not belong to conserved genomic neighborhoods, but several such putative operons were detected. Perhaps, the most notable one is the ESCRT neighborhood which includes genes coding for subunits of ESCRT I, II and III, and often, components of the ubiquitin system (*2*), suggesting an ancient link between the two systems that persists in eukaryotes (*35*). We predicted another operon that is conserved in most Asgard archaea and consists of genes encoding a LAMTOR1-like protein of the Roadblock superfamily, a Rab-like small GTPase, and a protein containing the DENN (differentially expressed in normal and neoplastic cells) domain that so far has been identified only in eukaryotes (Fig. 5b). Two proteins consisting of a DENN domain fused to longin are subunits of the folliculin (FLCN) complex that is conserved in eukaryotes. The FLCN complex is the sensor of amino acid starvation interacting with Rag GTPase and Ragulator lysosomal complex, and a key component of the mTORC1 pathway, the central regulator of cell growth in eukaryotes (*39*). Some Heimdallarchaea encode several proteins with the exact same domain organization as FLCN (Fig. 5b). Ragulator is a

complex that consists of 5 subunits, each containing the Roadblock domain. In Asgard archaea, however, the GTPase present in the operon is from a family that is distinct from the Rag GTPases, which interact with both FLCN and Ragulator complexes in eukaryotes, despite the fact that Rag family GTPases are abundant in Asgards (*33*) (Supplementary Table 7). Nevertheless, this conserved module of Asgard proteins is a strong candidate to function as a guanine nucleotide exchange factor for Rab and Rag GTPases, analogously to the eukaryotic FLCN. In eukaryotes, the DENN domain is present in many proteins with different domain architectures that interact with different partners and perform a variety of functions (*40*, *41*). The Asgard archaea also encode other DENN domain proteins, and the respective genes form expanded families of paralogs in Loki, Hel and Heimdall lineages, again, with domain architectures distinct from those in eukaryotes (Fig. 5b) (*42*).

Prompted by the identification of a FLCN-like complex, we searched for other components of the mTORC1 regulatory pathway in Asgard archaea. The GATOR1 complex that consists of three subunits, Depdc5, Nprl2, and Nprl3, is another amino acid starvation sensor that is involved in this pathway in eukaryotes (*43*). Nitrogen permease regulators 2 and 3 (NPRL2 and NPRL3) are homologous GATOR1 subunits that contain a longin domain and a small NPRL2-specific C-terminal domain (*43*). We identified a protein family with this domain organization in most Thor MAGs and a few Loki MAGs. Several other ESP asCOGs include proteins with high similarity to the longin domain of NPRL2. Additionally, we identified many fusions of the NPRL2-like longin domain with various domains related to prokaryotic two-component signal transduction system (Fig. 5c). Considering the absence of a homolog of phosphatidylinositol 3-kinase, the catalytic domain of the mTOR protein, it seems likely that, in Asgard archaea, the key growth regulation pathway remains centered at typical prokaryotic two-component signal transduction systems whereas at least some of the regulators and sensors in this pathway are “eukaryotic”. The abundance of NPRL2-like longin domains in Asgard archaea implies that the link between this domain and amino acid starvation regulation emerged at the onset of Asgard evolution if not earlier.

## Diverse metabolic repertoires, ancestral metabolism of Asgard archaea, and syntrophic evolution

Examination of the distribution of the asCOGs among the 12 Asgard archaeal phyla showed that the metabolic pathway repertoire was conserved among the MAGs of each phylum but differed between the phyla (Fig. 3a). Three distinct lifestyles were predicted by the asCOG analysis for different major branches of Asgard archaea, namely, anaerobic heterotrophy, facultative aerobic heterotrophy, and chemolithotrophy (Fig. 6, Supplementary Figure 9). For the last Asgard archaeal common ancestor (LAsCA), a mixotrophic life style, including both production and consumption of $H_2$, can be inferred

from parsimony considerations (Fig. 6, Supplementary Table 8; see Materials and Methods for further details). Loki-, Thor-, Hermod-, Baldr- and Borrarchaeota encode all enzymes for the complete (archaeal) Wood-Ljungdahl pathway (WLP) and are predicted to oxidize organic substrates, likely, by using the reverse WLP, given the lack of enzymes for oxidation of inorganic compounds (e.g., hydrogen, sulfur/sulfide and nitrogen/ammonia). The genomes of these five Asgard phyla encode homologues of membrane-bound respiratory $H_2$-evolving Group 4 [NiFe] hydrogenase (Supplementary Figure 6) and/or cytosolic cofactor-coupled bidirectional Group 3 [NiFe] hydrogenase (*44*) (Supplementary Figure 7). Phylogenetic analysis of both group 3 and group 4 [NiFe] hydrogenases showed that Asgard archaea form distinct clades well separated from the functionally characterized hydrogenases, hampering the prediction of their specific functions in Asgard archaea. The functionally characterized group 4 [NiFe] hydrogenases in the Thermococci are involved in the fermentation of organic substrates to $H_2$, acetate and carbon dioxide (*45*, *46*). The presence of group 3 [NiFe] hydrogenases suggests that these Asgard archaea cannot use $H_2$ as an electron donor because they lack the enzyme complex coupling $H_2$ oxidation to membrane potential generation. Thus, in these organisms, bifurcate electrons from $H_2$ are likely to be used to support the fermentation of organic substrates exclusively (*45–47*).

Both Wukongarchaeota genomes (As_075 and As_085) encode a bona fide membrane-bound Group 1k [NiFe] hydrogenase that could mediate hydrogenotrophic respiration using heterodisulfide as the terminal electron acceptor (*48*, *49*) (Fig. 6, Supplementary Figure 8, Supplementary Figure 10). The group 1k [NiFe] hydrogenase is exclusively found in methanogens of the order Methanosarcinales (Euryarchaoeta) (*50*), and it is the first discovery of the group 1 [NiFe] hydrogenase in the Asgard archaea. Wukongarchaeota also encode all enzymes for a complete WLP and a putative ADP-dependent acetyl-CoA synthetase for acetate synthesis. Unlike all other Asgard archaea, Wukongarchaeota lack genes for citrate cycle and beta-oxidation. Thus, Wukongarchaeota appear to be obligate chemolithotrophic acetogens. The genomes of Wukongarchaeota were discovered only in seawater of the euphotic zone of the Yap trench (0 m and 125 m). Dissolved $H_2$ concentration is known to be the highest in surface seawater, where the active microbial fermentation, compared to deep sea (*51*), could produce sufficient amounts of hydrogen for the growth of Wukongarchaeota. Hodarchaeota, Gerdarchaeota, Kariarchaeota, and Heimdallarchaeota share a common ancestor with Wukongarchaeota (Fig. 6). However, genome analysis implies different lifestyles for these organisms. Hod-, Gerd- and Kariarchaeota encode various electron transport chain components, including heme/copper-type cytochrome/quinol oxidase, nitrate reductase, and NADH dehydrogenase, most likely, allowing the use of oxygen and nitrate as electron acceptors during aerobic and anaerobic respiration, respectively (*44*). In addition, Hod-, Gerd- and Heimdallarchaeota encode phosphoadenosine phosphosulfate (PAPS) reductase and adenylylsulfate kinase for sulfate reduction, enabling the use of sulfate as electron acceptor during anaerobic respiration.

Gerd-, Heimdall-, and Hodarchaeota are only found in coastal and deep-sea sedimentary environments, whereas Kariarchaeota were found also in marine water. The versatile predicted metabolic capacities of these groups suggest that Hod-, Gerd- and Kariarchaeota might occupy both anoxic and oxic niches. In contrast, Heimdallarchaeota appear to be able to thrive only in anoxic environments.

In the widely considered syntrophy scenarios (*52*), eukaryogenesis has been proposed to involve metabolic symbiosis (syntrophy) between an archaeon and one or two bacterial partners which, in the original hydrogen/syntrophy hypothesis, were postulated to donate $H_2$ for methane or hydrogen sulfide production by the consortium (*53*, *54*). The syntrophic scenarios were boosted by the discovery of apparent syntrophy between *Candidatus* P. syntrophicum and Deltaproteobacteria which led to the proposal of the Etangle-Engulf-Endogenize ($E^3$) model of eukaryogenesis (*17*, *55*). Reconstruction of the Lokiarchaeon metabolism has suggested that this organism was hydrogen-dependent, in accord with the hydrogen-syntrophic scenarios (*54*). In contrast, subsequent analysis of the metabolic potentials of 4 Asgard phyla has led to the inference that these organisms were primarily organoheterotrophic and $H_2$-producing, the 'reverse flow model' model of protoeukaryote energy metabolism that involves electron or hydrogen flow from an Asgard archaeon to the alphaproteobacterial ancestor of mitochondria, in the opposite direction from that in the original hydrogen-syntrophy hypotheses (*44*). Here, we discovered a deeply branching Asgard group, Wukongarchaeota, which appears to include obligate hydrogenotrophic acetogens, suggesting the possibility of the LAsCA being a hydrogen-dependent autotroph (Fig. 6). This finding suggests that LAsCA both produced and consumed $H_2$. Thus, depending on the exact relationship between Asgard archaea and eukaryotes that remains to be elucidated, our findings could be compatible with different syntrophic scenarios that postulate $H_2$ transfer from bacteria to the archaeal symbiont or in the opposite direction.

## Conclusions

The Asgard archaea that were discovered only 5 years ago as a result of the painstaking assembly of several Loki Castle metagenomes have grown into a highly diverse archaeal superphylum. The most remarkable feature of the Asgards is their apparent evolutionary affinity with eukaryotes that has been buttressed by two independent lines of evidence: phylogenetic analysis of highly conserved genes and detection of multiple ESPs that are absent or far less common in other archaea. The 75 MAGs added here substantially expand the phylogenetic and metabolic diversity of the Asgard superphylum. The extended set of Asgard genomes provides for a phylogeny based on a far more representative species sampling than available previously and a substantially expanded ESP analysis employing powerful computational

methods. This extended analysis reveals 6 putative additional Asgard phyla but does not immediately clarify the Asgard-eukaryote relationship. Indeed, phylogenetic analysis of conserved genes in an expanded set of archaea, bacteria and eukaryotes yields conflicting 3D and 2D signals, with an unexpected preference for the 3D topology. Thus, the conclusion that eukaryotes emerged from within Asgard archaea, in particular, from the Heimdall lineage, appears to be premature. Further phylogenomic study with an even broader representation of diverse archaeal lineages as well as, possibly, even more sophisticated evolutionary models are required to clarify the relationships between archaea and eukaryotes.

Our analysis of Asgard genomes substantially expanded the set of ESPs encoded by this group of archaea and revealed numerous, complex domain architectures of these proteins. These results further emphasize the excess of ESPs in Asgards compared to other archaea and provide additional support to the conclusion that most of the Asgard ESPs are involved in membrane remodeling and intracellular trafficking. However, in parallel with the phylogenomic results, detailed analysis of the ESPs reveals a complex picture. Most of the multidomain Asgard ESPs possess domain architectures distinct from typical eukaryotic ones and some of these arrangements include signature prokaryotic domains, suggesting substantial functional differences from the respective eukaryotic systems. Furthermore, virtually all the ESPs show patchy distributions in Asgard and other archaea, indicative of a history of extensive HGT, gene losses and paralogous family expansion. All these findings seem to be best compatible with the model of a dispersed, dynamic archaeal 'eukaryome' (*30*) that widely spreads among archaea via HGT, so far reaching the highest ESP density in the Asgard archaea.

The results of this work cannot rule out the possibility of the emergence of eukaryotes from within the Asgard but seem to be better compatible with a different evolutionary scenario under which the conserved core of eukaryote genes involved in informational processes originates from an as yet unknown ancestor group that might be a deep archaeal branch or could lie outside the presently characterized archaeal diversity. These hypothetical ancestral forms might have accumulated components of the mobile archaeal 'eukaryome' to an even greater extent than the Asgards archaea, eventually, giving rise to eukaryote-like cells, likely, via a form of syntrophy with one or more bacterial partners. Combined genomic and (undoubtedly, far more challenging) biological study of diverse archaea is essential for further advancing our understanding of eukaryogenesis.

**Author contributions**

ML, EVK, KSM and YL initiated the study;; YL performed metagenomic assembly, binning, metabolism analysis; KSM, AN and YIW performed comparative genomic analysis; YL, KSM, YIW and WCH performed phylogenetic analysis; KSM and YIW constructed asCOGs; KSM, YIW, YL, ML, and EVK analyzed the data; YL, KSM, WCH, EVK, and ML wrote the manuscript that was read, edited and approved by all authors.

## Acknowledgements

ML and YL are supported by National Natural Science Foundation of China (Grant No. 91851105, 31970105 and 31700430), the Key Project of Department of Education of Guangdong Province (No.2017KZDXM071), and the Shenzhen Science and Technology Program (Grant no. JCYJ20170818091727570 and KQTD20180412181334790). KSM, YIW, SAS and EVK are supported by the Intramural Research Program of the National Institutes of Health of the USA (National Library of Medicine).

## Competing interests

The authors declare no competing interests

## References

1. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. **521**, 173–179 (2015).

2. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. B ckstr m, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A. Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*. **541**, 353–358 (2017).

3. F. MacLeod, G. S. Kindler, H. L. Wong, R. Chen, B. P. Burns, Asgard archaea: Diversity, function, and evolutionary implications in a range of microbiomes. *AIMS Microbiol*. **5**, 48–61 (2019).

4. M. Cai, Y. Liu, X. Yin, Z. Zhou, M. W. Friedrich, T. Richter-Heitmann, R. Nimzyk, A. Kulkarni, X. Wang, W. Li, J. Pan, Y. Yang, J.-D. Gu, M. Li, Diverse Asgard archaea including the novel phylum Gerdarchaeota participate in organic matter degradation. *Sci. China Life Sci.* **63**, 886–897 (2020).

5. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* (2019), doi:10.1038/s41559-019-1040-x.

6. T. A. Williams, P. G. Foster, C. J. Cox, T. M. Embley, An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*. **504**, 231–236 (2013).

7. C. J. Cox, P. G. Foster, R. P. Hirt, S. R. Harris, T. M. Embley, The archaebacterial origin of eukaryotes. *Proc Natl Acad Sci U S A*. **105**, 20356–20361 (2008).

8. N. Yutin, K. S. Makarova, S. L. Mekhedov, Y. I. Wolf, E. V. Koonin, The deep archaeal roots of eukaryotes. *Mol. Biol. Evol.* **25**, 1619–1630 (2008).

9. V. Da Cunha, M. Gaia, D. Gadelle, A. Nasir, P. Forterre, Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. *PLoS Genet.* **13**, e1006810 (2017).

10. V. D. Cunha, M. Gaia, A. Nasir, P. Forterre, Asgard archaea do not close the debate about the universal tree of life topology. *PLoS Genet.* **14**, e1007215 (2018).

11. P. Forterre, The universal tree of life: an update. *Front. Microbiol*. **6**, 717 (2015).

12. J. Lombard, P. López-García, D. Moreira, The early evolution of lipid membranes and the three domains of life. *Nat. Rev. Microbiol*. **10**, 507–515 (2012).

13. J. A. Burns, A. A. Pittis, E. Kim, Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nat. Ecol. Evol*. **6**, a016006 (2018).

14. C. Akıl, R. C. Robinson, Genomes of Asgard archaea encode profilins that regulate actin. *Nature*. **562**, 439–443 (2018).

15. C. Akıl, L. T. Tran, M. Orhant-Prioux, Y. Baskaran, E. Manser, L. Blanchoin, R. C. Robinson, Insights into the evolution of regulated actin dynamics via characterization of primitive gelsolin/cofilin proteins from Asgard archaea. *Proc Natl Acad Sci U S A*. **117**, 19904–19913 (2020).

16. Z. Lu, T. Fu, T. Li, Y. Liu, S. Zhang, J. Li, J. Dai, E. V. Koonin, G. Li, H. Chu, M. Li, Coevolution of Eukaryote-like Vps4 and ESCRT-III Subunits in the Asgard Archaea. *mBio*. **11**, e00417-20, /mbio/11/3/mBio.00417-20.atom (2020).

17. H. Imachi, M. K. Nobu, N. Nakahara, Y. Morono, M. Ogawara, Y. Takaki, Y. Takano, K. Uematsu, T. Ikuta, M. Ito, Y. Matsui, M. Miyazaki, K. Murata, Y. Saito, S. Sakai, C. Song, E. Tasumi, Y. Yamanaka, T. Yamaguchi, Y. Kamagata, H. Tamaki, K. Takai, Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature*. **577**, 519–525 (2020).

18. S. M. Karst, M. S. Dueholm, S. J. McIlroy, R. H. Kirkegaard, P. H. Nielsen, M. Albertsen, *Nat. Biotechnol.*, in press, doi:10.1038/nbt.4045.

19. R.-Y. Zhang, B. Zou, Y.-W. Yan, C. O. Jeon, M. Li, M. Cai, Z.-X. Quan, Design of targeted primers based on 16S rRNA sequences in meta-transcriptomic datasets and identification of a novel taxonomic group in the Asgard archaea. *BMC microbiology*. **20**, 25 (2020).

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

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

22. A. R. Nabhan, I. N. Sarkar, The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Brief Bioinform*. **13**, 122–134 (2012).

23. R. Gouy, D. Baurain, H. Philippe, Rooting the tree of life: the phylogenetic jury is still out. *Philos Trans R Soc Lond B Biol Sci*. **370**, 20140329 (2015).

24. J. Felsenstein, *Inferring Phylogenies* (Oxford University Press, Oxford, New York, 2nd Edition., 2003).

25. F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*. **311**, 1283–1287 (2006).

26. C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, D. W. Waite, W. B. Whitman, D. H. Parks, P. Hugenholtz, *bioRxiv*, in press, doi:10.1101/2020.03.01.972265.

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

28. A. B. Narrowe, A. Spang, C. W. Stairs, E. F. Caceres, B. J. Baker, C. S. Miller, T. J. G. Ettema, Complex Evolutionary History of Translation Elongation Factor 2 and Diphthamide Biosynthesis in Archaea and Parabasalids. *Genome Biol Evol*. **10**, 2380–2393 (2018).

29. L. Eme, A. Spang, J. Lombard, C. W. Stairs, T. J. G. Ettema, Archaea and the origin of eukaryotes. *Nat. Rev. Microbiol.* **15**, 711–723 (2017).

30. E. V. Koonin, N. Yutin, The dispersed archaeal eukaryome and the complex archaeal ancestor of eukaryotes. *Cold Spring Harb Perspect Biol*. **6**, a016188 (2014).

31. R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale, The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. **4**, 41 (2003).

32. E. V. Koonin, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, D. M. Krylov, K. S. Makarova, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, I. B. Rogozin, S. Smirnov, A. V. Sorokin, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, D. A. Natale, A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*. **5**, R7 (2004).

33. C. M. Klinger, A. Spang, J. B. Dacks, T. J. G. Ettema, Tracing the Archaeal Origins of Eukaryotic Membrane-Trafficking System Building Blocks. *Mol Biol Evol*. **33**, 1528–1541 (2016).

34. P. Puigbò, A. E. Lobkovsky, D. M. Kristensen, Y. I. Wolf, E. V. Koonin, Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol*. **12**, 66 (2014).

35. L. Christ, C. Raiborg, E. M. Wenzel, C. Campsteijn, H. Stenmark, Cellular Functions and Molecular Mechanisms of the ESCRT Membrane-Scission Machinery. *Trends Biochem Sci*. **42**, 42–56 (2017).

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

37. K. M. K. Kibria, J. Ferdous, R. Sardar, A. Panda, D. Gupta, A. Mohmmed, P. Malhotra, A genome-wide analysis of coatomer protein (COP) subunits of apicomplexan parasites and their evolutionary relationships. *BMC Genomics*. **20**, 98 (2019).

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

39. R. E. Lawrence, S. A. Fromm, Y. Fu, A. L. Yokom, D. J. Kim, A. M. Thelen, L. N. Young, C.-Y. Lim, A. J. Samelson, J. H. Hurley, R. Zoncu, Structural mechanism of a Rag

538 GTPase activation checkpoint by the lysosomal folliculin complex. *Science*. **366**, 971–977
539 (2019).

540 40. M.-Y. Su, S. A. Fromm, R. Zoncu, J. H. Hurley, Structure of the C9orf72 ARF GAP
541 complex that is haploinsufficient in ALS and FTD. *Nature*. **585**, 251–255 (2020).

542 41. N. de Martín Garrido, C. H. S. Aylett, Nutrient Signaling and Lysosome Positioning
543 Crosstalk Through a Multifunctional Protein, Folliculin. *Front Cell Dev Biol*. **8**, 108 (2020).

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

546 43. K. Shen, R. K. Huang, E. J. Brignole, K. J. Condon, M. L. Valenstein, L. Chantranupong,
547 A. Bomaliyamu, A. Choe, C. Hong, Z. Yu, D. M. Sabatini, Architecture of the human GATOR1
548 and GATOR1-Rag GTPases complexes. *Nature*. **556**, 64–69 (2018).

549 44. A. Spang, C. W. Stairs, N. Dombrowski, L. Eme, J. Lombard, E. F. Caceres, C.
550 Greening, B. J. Baker, T. J. G. Ettema, Proposal of the reverse flow model for the origin of the
551 eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nat. Microbiol*. **4**,
552 1138–1148 (2019).

553 45. H. Yu, C.-H. Wu, G. J. Schut, D. K. Haja, G. Zhao, J. W. Peters, M. W. W. Adams, H.
554 Li, Structure of an Ancient Respiratory System. *Cell*. **173**, 1636-1649.e16 (2018).

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

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

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

563 49. R. K. Thauer, A.-K. Kaster, M. Goenrich, M. Schick, T. Hiromoto, S. Shima,
564 Hydrogenases from methanogenic archaea, nickel, a novel cofactor, and H2 storage. *Annu Rev*
565 *Biochem*. **79**, 507–536 (2010).

566 50. D. Søndergaard, C. N. S. Pedersen, C. Greening, HydDB: A web tool for hydrogenase
567 classification and analysis. *Scientific Reports*. **6**, 34212 (2016).

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

570 52. P. López-García, D. Moreira, The Syntrophy hypothesis for the origin of eukaryotes
571 revisited. *Nat Microbiol*. **5**, 655–667 (2020).

572 53. W. Martin, M. Müller, The hydrogen hypothesis for the first eukaryote. *Nature*. **392**, 37–
573 41 (1998).

574 54. D. Moreira, P. López-García, Symbiosis Between Methanogenic Archaea and δ-
575 Proteobacteria as the Origin of Eukaryotes: The Syntrophic Hypothesis. *J Mol Evol*. **47**, 517–530
576 (1998).

577 55. P. López-García, D. Moreira, Cultured Asgard Archaea Shed Light on Eukaryogenesis.
578 *Cell*. **181**, 232–235 (2020).

579

580

581

582 .

583

584

585 **Figures**

586

587 **Figure 1. Gene commonality plots for Asgard archaea and the TACK superphylum.**

588 The gene commonality plot showing the number of asCOGs in log scale (Y-axis) that include the given
589 fraction of analyzed genomes (X-axis). The Asgard plot is compared with the TACK superphylum plot
590 based on assignment of TACK genomes to arCOGs.

591

592 **Figure 2. Phylogenetic analysis of Asgard archaea and their relationships with eukaryotes.**

593 **(a)** Maximum likelihood tree, inferred with IQ-tree and LG+F+R10 model, constructed from
594 concatenated alignments of the protein sequences from 209 core Asgard Clusters of Orthologs (asCOGs).
595 Only the 12 phylum-level clades are shown, with species within each clade collapsed. See supplementary
596 Methods and Supplementary Table 5 for details.

597 **(b)** Maximum likelihood tree, inferred with IQ-tree and SYM+R8 model, based on 16S rRNA gene
598 sequences. Red stars in **(b)** denote MAGs reconstructed in the current study.

599 (**c**) Phylogenetic tree of bacteria, archaea and eukaryotes, inferred with IQ-tree under LG+R10 model,
600 constructed from concatenated alignments of the protein sequences of 30 universally conserved genes
601 (see Material and Methods for details). The tree shows the relationships between the major clades.

602 The trees are unrooted and are shown in a pseudorooted form for visualization purposes only. The actual
603 trees and alignments are in Additional data file 2 and list of the trees are provided in the Supplementary
604 Table 4 and 5.

605

606 **Figure 3. Phyletic patterns of asCOGs and functional distribution of Asgard core genes.**

607 **(a)** Classical Multidimensional Scaling analysis of binary presence-absence phyletic patterns for 13,939
608 asCOGs that are represented in at least two genomes (see Material and Methods for details).

609 **(b)** Functional breakdown of Asgard core genes (378 asCOGs) compared with TACK superphylum core
610 genes (489 arCOGs). The values were normalized as described in Materials and Methods. Functional
611 classes of genes: J, Translation, ribosomal structure and biogenesis; K, Transcription; L, Replication,

recombination and repair; D, Cell cycle control, cell division, chromosome partitioning; V, Defense mechanisms; T, Signal transduction mechanisms; M, Cell wall/membrane/envelope biogenesis; N, Cell motility; U, Intracellular trafficking, secretion, and vesicular transport; O, Posttranslational modification, protein turnover, chaperones; X, Mobilome: prophages, transposons; C, Energy production and conversion; G, Carbohydrate transport and metabolism; E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; H, Coenzyme transport and metabolism; I, Lipid transport and metabolism; P, Inorganic ion transport and metabolism; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function prediction only; S, Function unknown;

**Figure 4. Phyletic patterns of Eukaryotic Signature Proteins (ESPs) encoded in Asgard genomes.**

All 505 ESP asCOGs are grouped by distance between binary presence-absence phyletic patterns. The most highly conserved ESP asCOGs are shown within the red rectangle. Below the plot of the number of ESP domains in each genome is shown. For details, see Supplementary Table 7.

**Figure 5. Domain architectures of selected Asgard ESPs.**

**(a) ESPs with unique domain architectures.** The schematic of each multidomain protein is roughly proportional to the respective protein length. The identified domains are shown inside the arrows approximately according to their location and are briefly annotated. Homologous domains are shown by the same color or pattern.

**(b) DENN domain proteins in Asgards.** Upper part of the figure above the dashed line shows putative operon encoding DENN domain proteins. Genes are shown by block arrows with the length proportional to the size of the corresponding protein. For each protein, the nucleotide contig or genome partition accession number, Asgard genome ID and lineage are indicated. The part of the figure below the dashed line shows domain organization of diverse proteins containing DENN domain. Homologous domains are shown by the same color. The inset on the right side explains identity of the domains

**(c) NPRL2-like proteins in Asgards.** Designations are the same as in Figure 4a.

**Figure 6. Reconstruction and evolution of the key metabolic processes in Asgard archaea.**

The schematic phylogeny of Asgard archaea is from Figure 1a.

LAsCA, Last Asgard Common Ancestor; WLP, Wood-Ljungdahl pathway.

642 **Supplementary Materials**

643 Taxonomic description

644 Materials and Methods

645 Supplementary References

646 Tables S1-S8

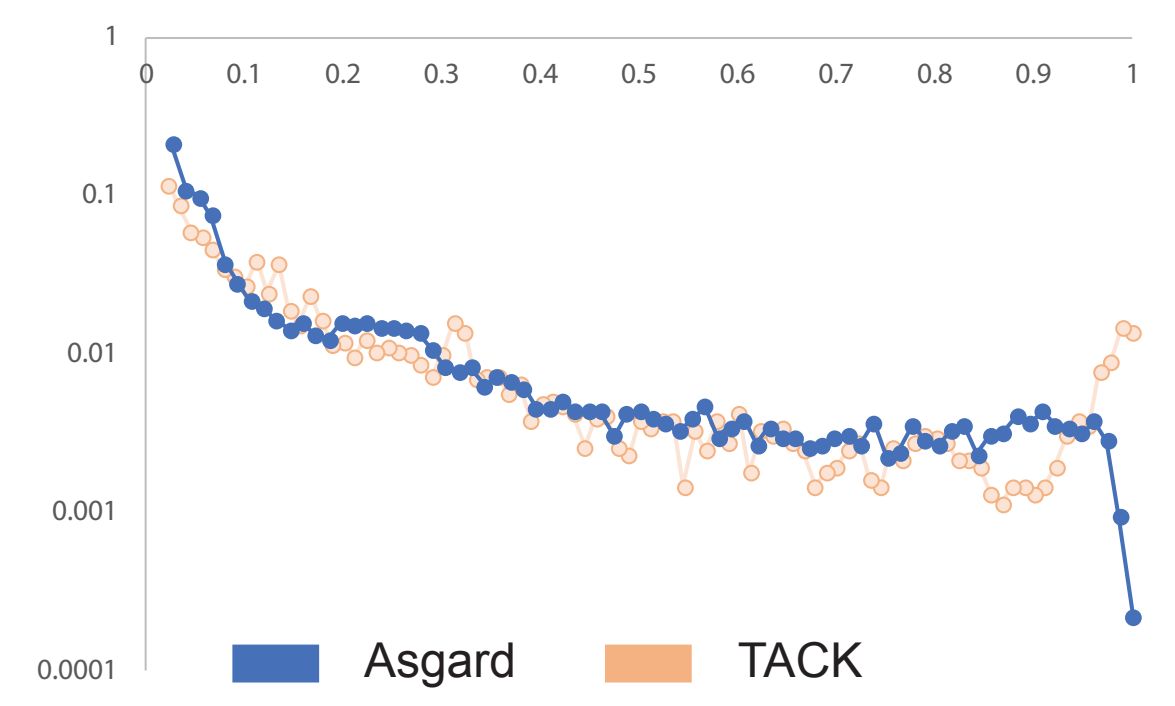
647 Figures S1-S10

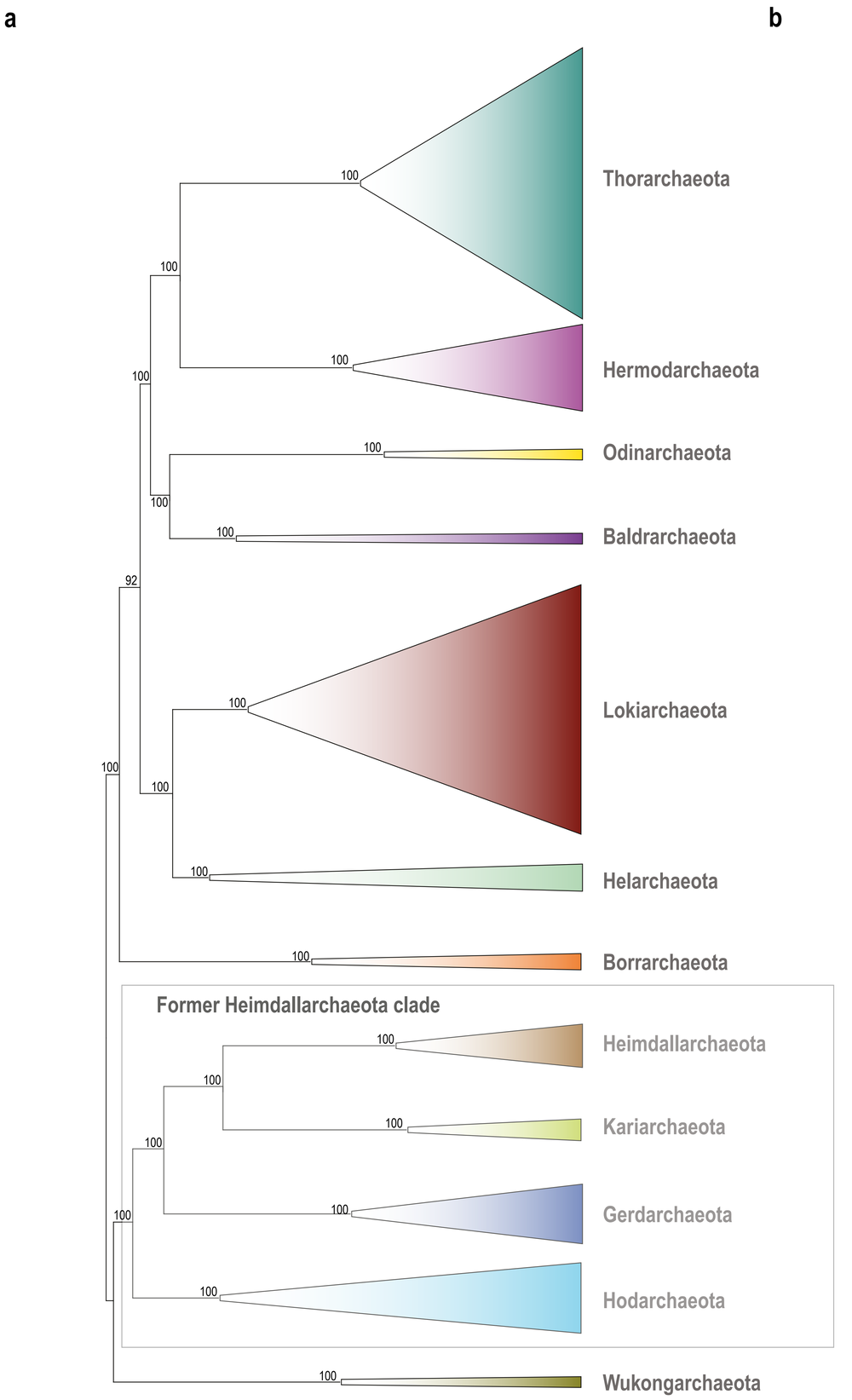
648 Additional Data files are available at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/asgard20/

649 Additional data file 1 (Additional_data_file_1.tgz): Complete asCOG data archive

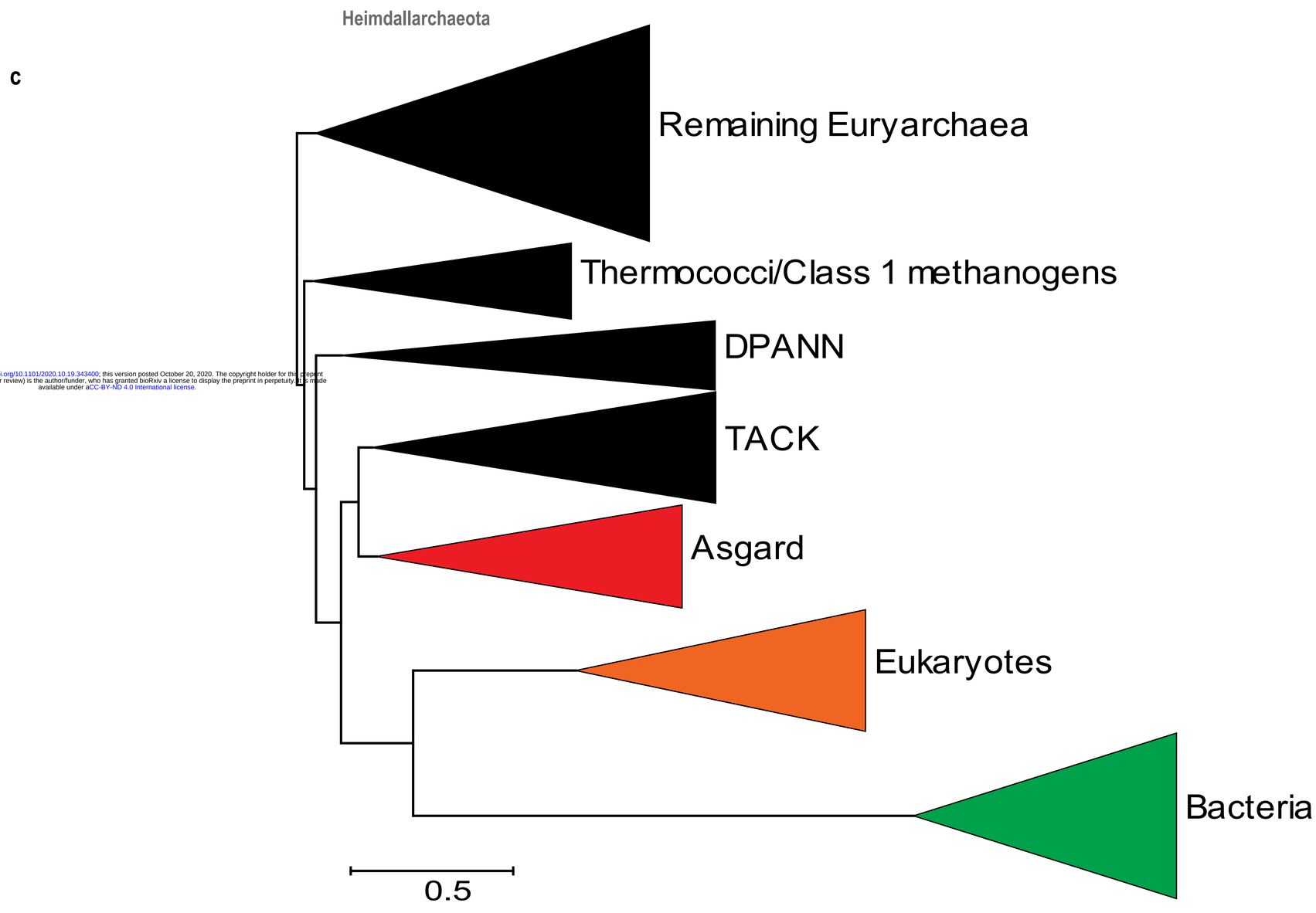
650 Additional data file 2 (Additional_data_file_2.tgz): Phylogenetic trees and alignments archive

651

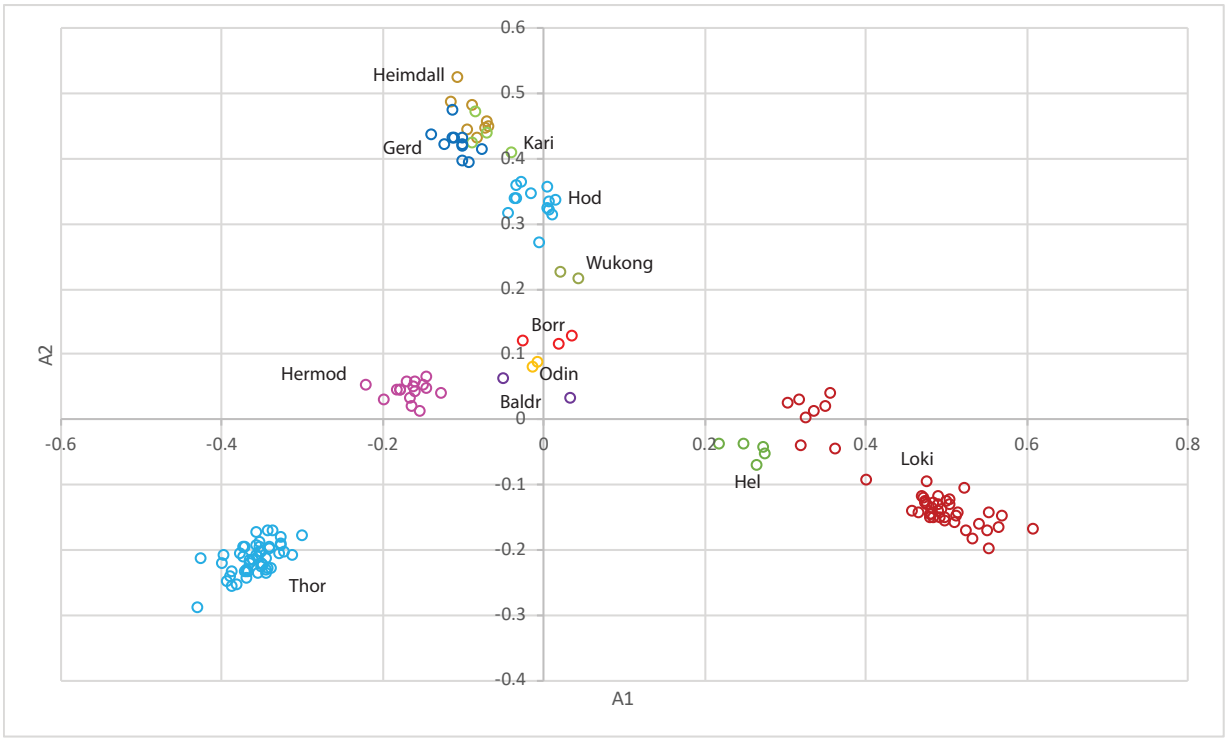

1
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.1
0.01
0.001
0.0001
Asgard
TACK



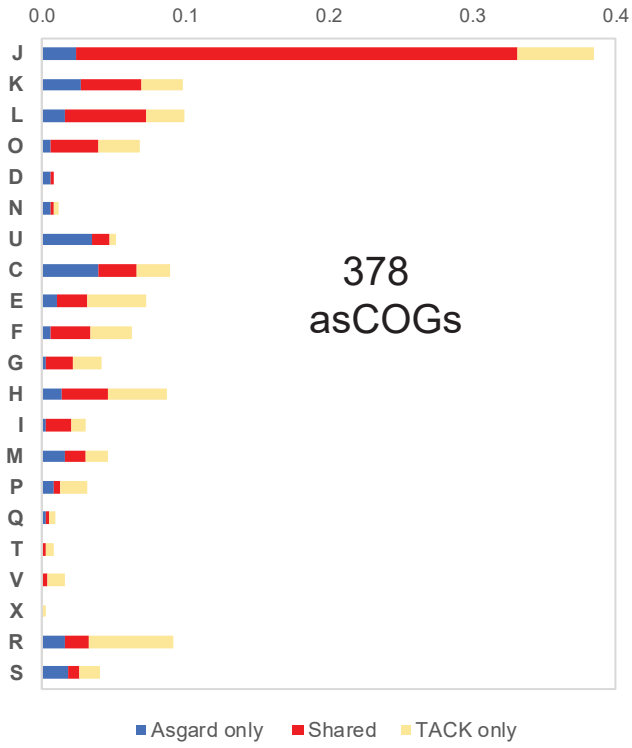


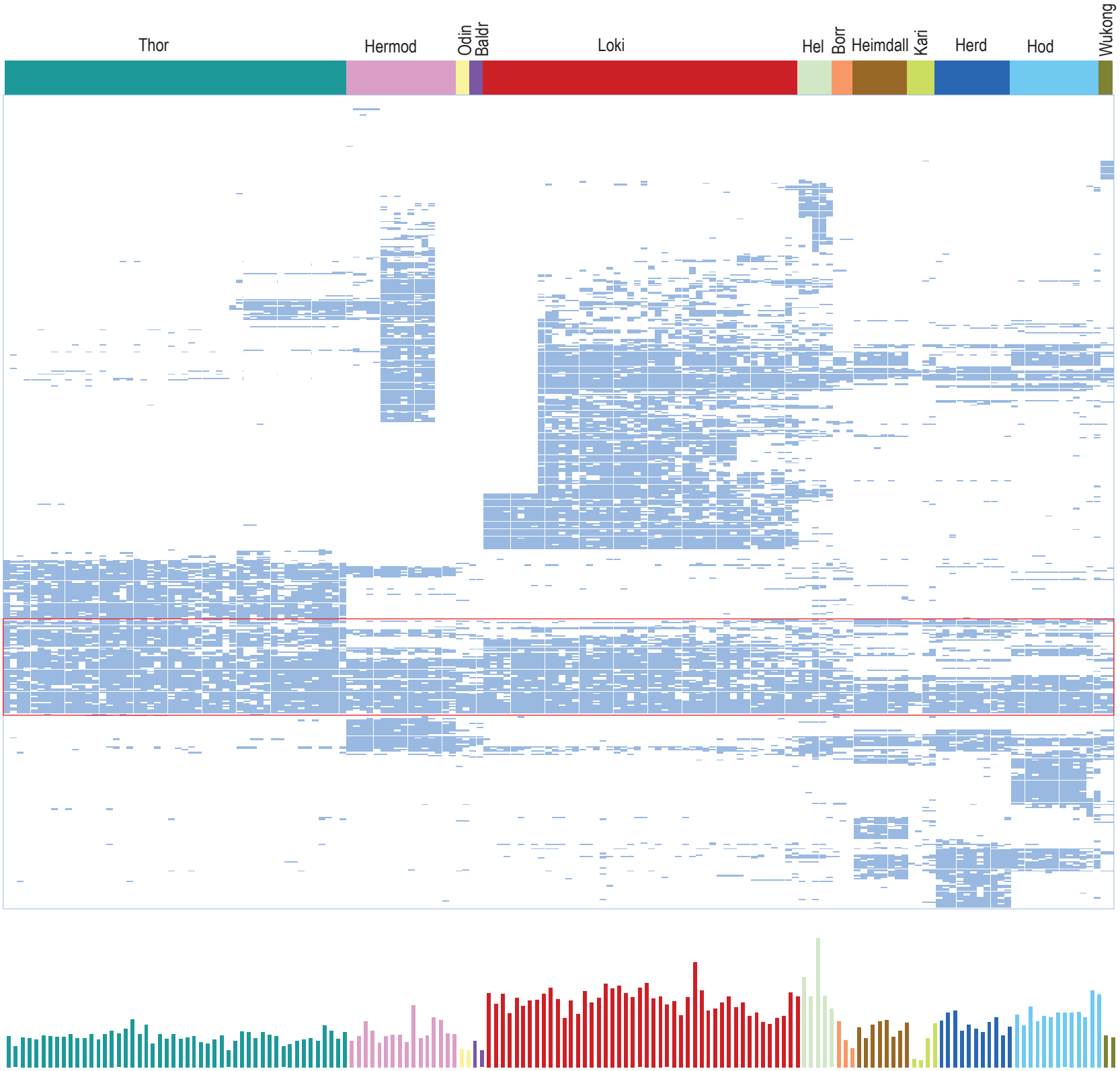


A

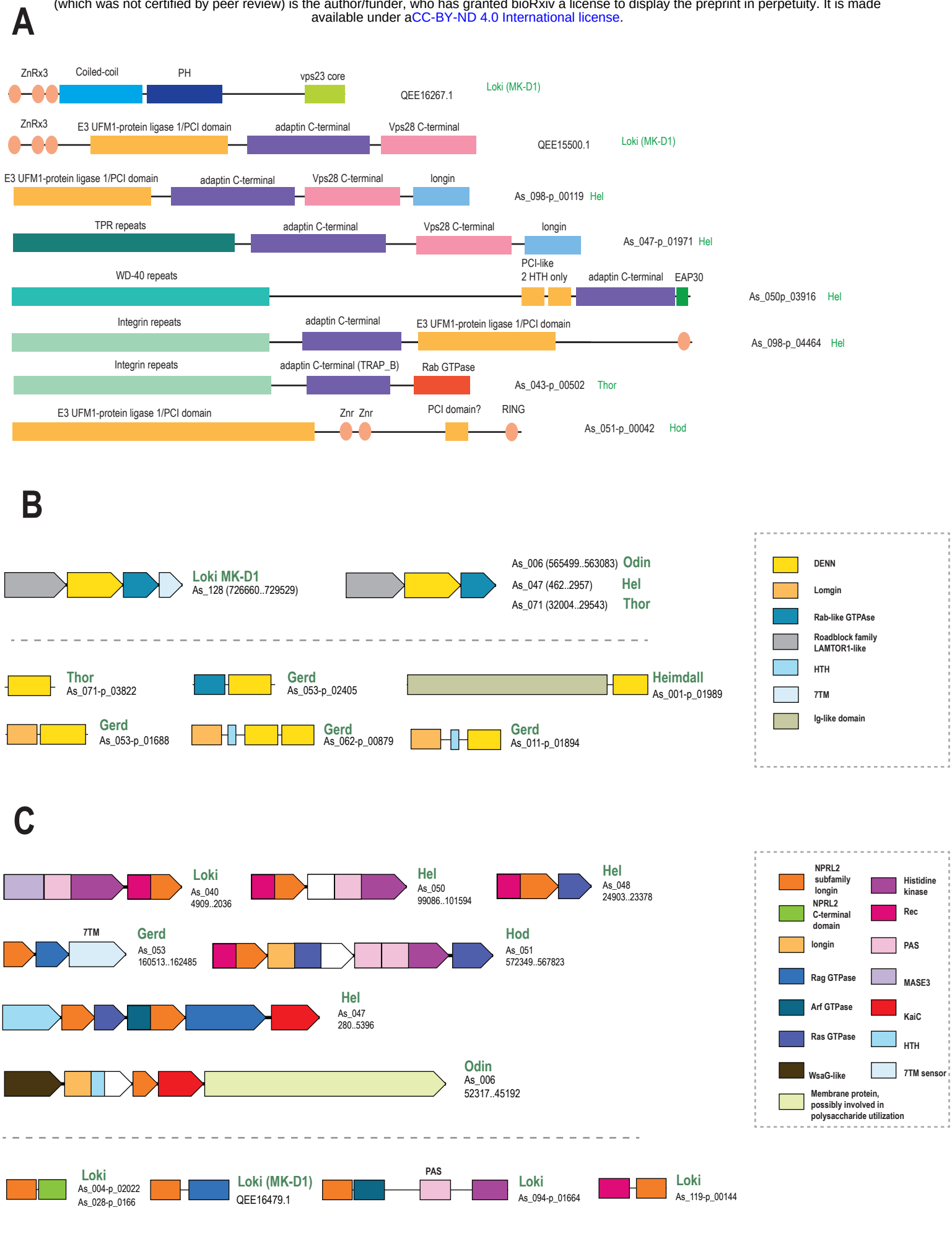


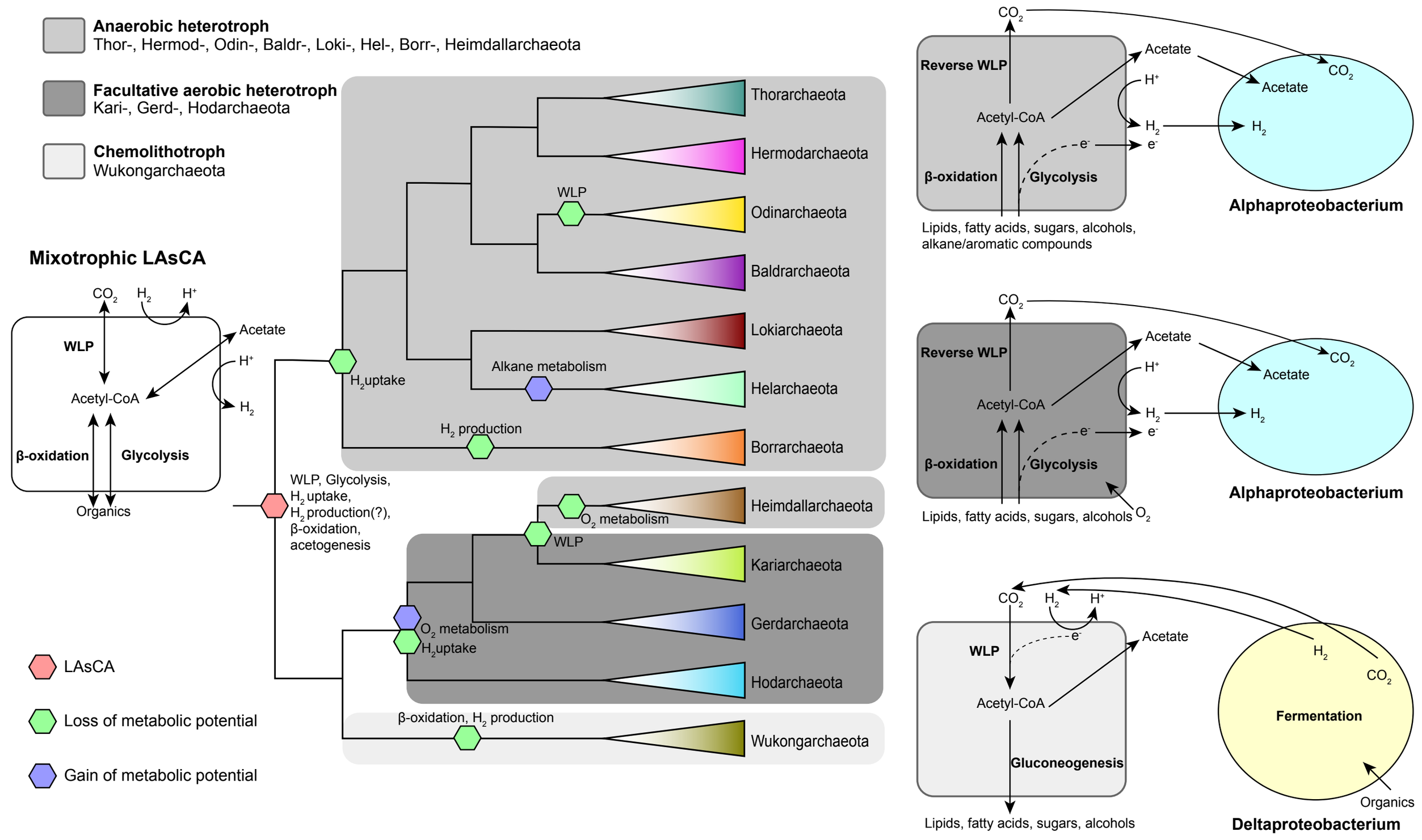

B






ESCRT I, II, III;
actin, gelsolins and profilins;
ubiquitin-system components E1,E2,E3;




Anaerobic heterotroph
Thor-, Hermod-, Odin-, Baldr-, Loki-, Hel-, Borr-, Heimdallarchaeota
Facultative aerobic heterotroph
Kari-, Gerd-, Hodarchaeota
Chemolithotroph
Wukongarchaeota
Mixotrophic LAsCA
CO2
H2
H+
WLP
Acetate
Acetyl-CoA
β-oxidation
Glycolysis
Organics
Thorarchaeota
Hermodarchaeota
WLP
Odinarchaeota
Baldrarchaeota
Lokiarchaeota
Alkane metabolism
Helarchaeota
H2uptake
H2 production
Borrarchaeota
WLP, Glycolysis,
H2 uptake,
H2 production(?),
β-oxidation,
acetogenesis
Heimdallarchaeota
O2 metabolism
WLP
Kariarchaeota
Gerdarchaeota
O2 metabolism
H2uptake
Hodarchaeota
β-oxidation, H2 production
Wukongarchaeota
LAsCA
Loss of metabolic potential
Gain of metabolic potential
Reverse WLP
Acetyl-CoA
β-oxidation
Glycolysis
e-
Lipids, fatty acids, sugars, alcohols,
alkane/aromatic compounds
Alphaproteobacterium
Lipids, fatty acids, sugars, alcohols
O2
Alphaproteobacterium
WLP
Acetyl-CoA
Gluconeogenesis
Fermentation
Organics
Lipids, fatty acids, sugars, alcohols
Deltaproteobacterium

# Expanding diversity of Asgard archaea and the elusive ancestry of eukaryotes

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

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

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

3 State Key Laboratory Breeding Base of Marine Genetic Resources, Key Laboratory of Marine Genetic Resources, Fujian Key Laboratory of Marine Genetic Resources, Third Institute of Oceanography, State Oceanic Administration, Xiamen 361005, P. R. China

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

† These authors contributed equally to this work.

*Authors for correspondence: koonin@ncbi.nlm.nih.gov or limeng848@szu.edu.cn

17 **Supplementary Information**

18 Taxonomic Description of new taxa……..…1

19 Materials and Methods……..……..9

20 Supplementary references…...……15

21 Supplementary Figures……...19

22 Supplementary Tables……...…….…….21

23

24

## Taxonomic Description of new taxa

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

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

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

**'*Candidatus* Hodarchaeum mangrovi'** (man.gro'vi N.L. fem. n. of a mangrove, referring to the isolation of the type material from mangrove soil). Type material is the genome designated as As_027 (FT2_5_011) representing '*Candidatus* Hodarchaeum mangrovi'. The genome "As_027" represents a MAG consisting of 4.01 Mbps in 348 contigs with an estimated completeness of 93.61%, an estimated contamination of 0.93%, a 23S rRNA gene and 14 tRNAs. The MAG recovered from mangrove sediment metagenomes (Futian Nature Reserve, China), with an estimated depth of coverage of 17.9, has a GC content of 32.9%.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## Materials and Methods

### Sampling collections and DNA sequencing

YT samples were obtained from the Rongcheng Nation Swan Nature Reserve (Rongcheng, China) in November 2018. The sediment cores were collected using columnar samplers at depth intervals of 0–2, 21–26, and 36–41 cm at a seagrass meadow and a non-seagrass–covered site nearby. After collection, bulk sediments were immediately sealed in plastic bags, placed in a pre-cooled icebox, and transported to laboratory within 4 hours. For each sample, DNA was extracted from 10 g sediment using PowerSoil DNA Isolation kit (QIAGEN, Germany), according to the manufacturer's protocol. Following extraction, nucleic acids were sequenced using Illumina HiSeq2500 (Illumina, USA) PE150 by Novogene (Nanjing, China).

MP5 samples were obtained from Mai Po Nature Reserve (Hong Kong, China) in September 2014. Three subsurface sediment samples were collected from a site covering with mangrove forest at depth intervals of 0-2, 10-15 and 20-25 cm. Two subsurface sediment samples were taken at an intertidal mudflat with depths of 0-5 and 13-16 cm. Samples were transported back to laboratory as described for YT metagenomes. DNA was extracted from 5g wet sediment per sample using the PowerSoil DNA Isolation Kit (MO BIO, USA) following the manufacturer's protocol. Metagenomic sequencing data were generated using Illumina HiSeq2500 (Illumina, USA) PE150 by Novogene (Tianjin, China).

FT samples were taken from Futian Nature Reserve (Shenzhen, Guangdong, China) in April 2017. Sediment samples were collected as described for YT samples at depth intervals of 0-2, 6-8, 12-14, 20-22, and 28-30 cm. DNA was extracted from 5g wet sediment per sample using DNeasy PowerSoil kit (Qiagen, Germany) as per manufacturer's instructions. Nucleic acids were sequenced using Illumina HiSeq2000 (Illumina, USA) PE150 by Novogene (Tianjin, China).

The surface sediment sample of the CJE metagenome was collected from Changjiang estuary during a cruise in August 2016. The sample was grabbed from the water bed, sealed immediately in a 50 ml tubes and stored in liquid nitrogen onboard. After transportation to laboratory, 10g wet sediments were used for DNA extraction as per manufacturer's protocol. Nucleic acids were sequenced using Illumina HiSeq2000 (Illumina, USA) PE150 by Novogene (Tianjin, China).

An oil sand sample was collected from Shengli Oilfield (Shandong, China) into bottles and they were transported to laboratory where they were stored at 4 ℃. The sample was used as inoculum to perform enrichment with anaerobic medium in vials as described elsewhere (*1*). After 253d of enrichment, the genomic DNA was extracted as described elsewhere (*2*). Nucleic acids were sequenced Illumina HiSeq2000 (Illumina, USA) PE150 by Novogene (Tianjin, China).

The seawater samples of Yap metagenomes were collected at Yap trench region by CTD SBE911plus (Sea-Bird Electronics, USA) during the 37$^{th}$ Dayang cruise in 2016. 8L of seawater per sample was filtered through a 0.22 μm-mesh membrane filter immediately after recovery onboard. The membrane was then cut into ~0.2 cm$^2$ pieces with a flame-sterilized scissors and added to a PowerBead Tube (MO BIO, USA) and the subsequent steps were implemented according to the manufacturer's protocol to extract DNA. The DNA per sample was amplified in five separative reactions using REPLI-g Single Cell Kits (Qiagen, Germany) following the manufacturer's protocol, given to the challenging nature of sample retrieval and DNA recovery. The products were pooled together and purified using QIAamp DNA Mini

Kit (Qiagen, Germany) according to the manufacturer's recommendations. Parallel blank controls were set for sampling, DNA extraction and ampliation with 0.22 μm-mesh membrane filtering Milli-Q water (18.2 MΩ; Millipore, USA). Nucleic acids were sequenced using HiSeq X Ten (Illumina, USA) PE150.

**Metagenomic assembly, binning and gene calling**

*FT, MP5 and YT metagenomes*. These three sets of metagenomes were assembled and binned using the same method. Raw shotgun metagenomic sequencing reads were trimmed with "read_qc" module from metaWRAP (v.1.1) (*3*). All clean reads from the same set were pooled together prior to *de novo* assemble to one co-assembly. Clean reads were sent out to MEGAHIT (v1.1.2) with flag "--presets meta-large" for co-assembling job (*4*). Sequencing coverage was determined for each assembled scaffold by mapping reads from each sample to the co-assembly using Bowtie2 (*5*). The binning analysis was carried out 8 times with 8 different combinations of specificity and sensitivity parameters using MetaBAT2(52) ("--maxP 60 or 95" AND "--minS 60 or 95" AND "--maxEdges 200 or 500") on the assembly with a minimum length of 2000 bp (*6*). DAS Tool (v1.0) was used as a dereplication and aggregation strategy on those eight binning results to construct accurate bins (*7*). Manual curation was used for reducing the genome contamination based on differential coverages, GC contents, and the presence of duplicate genes.

The depth coverage and N50 statistics of 38 Asgard MAGs recovered from YT metagenomes range from 7.72 to 298.86 (median: 21.77) and from 6658 to 381755 bps (median: 26337.5 bps) respectively; for 13 Asgard MAGs recovered from FT metagenomes, the values range from 6.98 to 33.22 (median 16.89) and from 3889 to 8957 bps (median: 5000 bps), respectively; and for 11 Asgard MAGs recovered from MP5 metagenomes, the values range from 7.06 to 54.42 (median: 10) and from 3898 to 19362 bps (median: 7581 bps) respectively. (**Supplementary Figure 2**).

*CJE metagenome*. Raw metagenomic shotgun sequencing reads were trimmed using Sickle (https://github.com/najoshi/sickle) with default settings. The trimmed reads were de novo assembled using IDBA-UD (v 1.1.1) with the parameters: "-mink 65, -maxk 145, -step 10" (*8*). Sequencing coverage was determined as described above. The binning analysis were performed with MetaBAT2 12 times, with 12 combinations of specificity and sensitivity parameters (--m 1500, 2000, or 2500" AND "--maxP 85 or 90" AND "--minS 80, 85 and 90") for further refinement (*6*). All binning results were merged and refined using DAS Tool (v1.0) (*7*).

2 Asgard MAGs recovered from CJE metagenome have a depth coverage of 19.16 and 20.56, and a N50 statistics of 8740 and 5246 bps (**Supplementary Figure 2**).

*J65 metagenome*. Raw metagenomic shotgun sequencing reads were trimmed with Trimmomatic (v0.38) (*9*). The clean reads were then fed to SPAdes (v 3.12.0) for *de novo* assembly with the parameters: "--meta -k 21, 33, 55, 77" (*10*). Sequencing coverage was determined using BBMap (v 38.24) toolkit with the parameters: "bbmap.sh minid=0.99" (https://github.com/BioInfoTools/BBMap). MetaBAT2 (v 2.12.1) was used to perform binning analysis with the parameter: "-m 2000" (*6*).

One Asgard MAG recovered from J65 metagenome had a depth coverage of 14.52 and a N50 statistics of 10460 bps (**Supplementary Figure 2**).

*Yap metagenome*. For each Yap metagenome, raw metagenomic shotgun sequencing reads were trimmed with Trimmomatic (v0.38) (*9*). Assembly and binning analysis were performed as described for CJE metagenome for each Yap metagenome.

The depth coverage and N50 statistics of 24 Asgard MAGs recovered from Yap metagenomes range from 7.78 to 82.33 (median: 13.91) and from 5097 to 889102 bps (median: 18155.5 bps) respectively. (**Supplementary Figure 2**)

A total of 89 Asgard MAGs were reconstructed in this work. Additional 95 Asgard MAGs were downloaded from public databases (e.g. NCBI FTP site). For all 184 genomes, a uniform gene calling protocol was applied. Specifically, the completeness, contamination, and strain heterogeneity of the genomes were estimated by using CheckM (v.1.0.12) (*11*) and DAS Tool under the taxonomic scope of domain (i.e., Bacteria and Archaea). Protein-coding genes were predicted using Prodigal (v 2.6.3) (*12*) embedded in Prokka (v 1.13) (*13*). Transfer RNAs (tRNAs) were identified with tRNAscan-SE (v1.23) using the archaeal tRNA model (*14*). After quality screening, further analysis focused on 162 high quality Asgard MAGs.

**Genome set**

The Asgard archaea genome set analyzed in this work consisted of 161 Asgard MAGs and one complete Asgard genome (**Supplementary Table 1**). For comparison, selected representative genomes of archaea (296), bacteria (100) and eukaryotes (76) were downloaded from Refseq and Genbank using the NCBI FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/).

**Average amino-acid identity**

The average amino-acid identity (AAI) across TACK archaeal reference genomes and the 184 Asgard genomes was calculated using compareM (v0.0.23) with the "aai_wf" at default settings (https://github.com/dparks1134/CompareM).

**asCOGs construction**

Initial clustering of 250,634 proteins encoded in 76 Asgard MAGs was performed using two approaches: first, footprints of arCOG profiles were obtained by running PSI-BLAST (*15*), initiated with arCOG alignments, against the set of predicted Asgard proteins. The footprint sequences were extracted and clustered according to the arCOG best hit. The remaining protein sequences (both full-length proteins and the sequence fragments outside of the footprints, if longer than 60 aa) were clustered using MMseqs2 (*16*), with similarity threshold of 0.5. Sequences within clusters were aligned using MUSCLE (*17*); the resulting alignments were passed through several rounds of merging and splitting. The merging phase involved comparing alignments to each other using HHSEARCH (*18*), finding full-length cluster-to-cluster matches, merging the sequence sets and re-aligning the new clusters. The splitting phase consisted of the construction of an approximate phylogenetic tree of the sequences using FastTree (*19*) (gamma-distributed site rates, WAG evolutionary model) with balanced mid-point tree rooting, identification of subtrees maximizing the fraction of species (MAGs) representation and minimizing the number of paralogs, and pruning such subtrees as separate clusters of putative orthologs. Clusters, derived from arCOGs, were prohibited from merging across distinct arCOGs to prevent distant paralogs from forming mixed clusters.

## Phylogenetic analysis

*16S rRNA gene phylogenetic analysis*. 16S rRNA gene sequences were identified in 73 genomes of Asgard archaea (26 generated in this study, 47 from public database) using Barrnap (v 0.9) with the "--kingdom arc" option (https://github.com/tseemann/barrnap). These sequences were combined with 46 published 16S rRNA sequences of Asgard archaea, to assess the novelty of the sequences obtained in this work. The novelty of 16S rRNA gene sequences was measured in terms of their sequence identity to previously identified Asgard archaeal 16S rRNA gene and phylogenetic relationships. Specifically, the pairwise sequence identity of two Asgard archaeal 16S rRNA gene sequences (>1300bp) was obtained by first globally aligning the sequences with "Stretcher" in EMBOSS package and then calculating the percent identity excluding gaps (*20*). The Asgard archaeal 16S rRNA gene sequences were aligned with 311 reference sequences from Euryarchaeota (n=231), DPANN (n=22), Korarchaeota (n=1), Crenarchaeota (n=41), Bathyarchaeota (n=2), Thaumarchaeota (n=13) and Aigarchaeota (n=1) using mafft-linsi (v7.471) (*21*) and trimmed with BMGE (v1.12) (settings: -m DNAPAM250:4, -g 0.5) (*22*).The alignment was used for phylogenetic inference with IQ-Tree (v2.0.6) based on SYM+R8 (selected by ModelFinder) to generate a maximum-likelihood tree (*23*).

*Asgard phylogeny.* A set of asCOGs that were considered most suitable as phylogenetic markers for Asgard archaea was selected using the preliminary classification of the 76 genomes in AsCOGs into previously described lineages: Loki, Thor, Odin, Hel and Heimdall. The following criteria were adopted: the asCOG have to be i) present in at least half of the genomes in all lineages, ii) present in at least 75% among the 76 genomes, iii) the mean number of paralogs per genome not to exceed 1.25. For the 209 asCOGs matching these criteria, the corresponding protein sequences were obtained from the extended set of 162 MAGs and aligned using MUSCLE (*17*); the 'index' paralog to include in the phylogeny was selected for each MAG based on the similarity to the alignment consensus. Alignments were trimmed to exclude columns containing more than 2/3 gap characters and with homogeneity below 0.05 (homogeneity is calculated from the score of the consensus amino acid against the alignment column, compared to the score of the perfect match (*24*) and concatenated, resulting in an alignment of 50,706 characters from 162 sequences (one sequence per MAG). Phylogeny was reconstructed using FastTree (gamma-distributed site rates, WAG evolutionary model) (*19*) and IQ-Tree (LG+F+R10 model, selected by ModelFinder) (*23*), producing very similar tree topologies.

*Tree of life.* To elucidate the relationships between the Asgard archaea and other major clades of archaea, bacteria and eukaryotes, 30 families of conserved proteins were selected that appear to have evolved mostly vertically (25, 26). The set of 162 MAGs of Asgard archaea was supplemented with 66 TACK archaea and 220 non-TACK archaea (the former, having been described as the closest archaeal relatives of the Asgard archaea (*27*), were sampled more densely), 98 bacteria and 72 eukaryotes. Prokaryotic genomes were sampled from the set of completely sequenced genomes to represent the maximum diversity within their respective clades (briefly, all proteins, encoded by genomes within a groups were clustered at 75% identity level; distances between genomes were estimated from the number of shared proteins within these clusters; UPGMA trees were reconstructed from the distances, and genome sets maximizing the total tree branch length were selected to represent the groups). The set of eukaryotic genomes was manually selected to represent the maximum possible variety of eukaryotic taxa. Genomes, in which more than four markers were missing, were excluded from the bacterial and eukaryotic sets. When multiple paralogs of a marker were present in a genome, preliminary phylogenetic trees were

constructed from protein sequence alignments, and paralogs with the shortest branches were selected to represent the corresponding genomes in the set. Sequences were aligned using MUSCLE (*17*), and alignment columns, containing more than 2/3 gap characters or with alignment column homogeneity below 0.05 were removed (*24*). The resulting concatenated alignments of the 30 markers consisted of 7411 sites. Phylogeny was reconstructed using FastTree (gamma-distributed site rates, WAG evolutionary model) (*19*) and IQ-Tree (*23*) with three models: LG+R10, selected by IQ-tree ModelFinder as the best fit, GTR20+F+R10 (following the suggestion of Williams et al. (*28*) to use GTR, we let IQ-tree to select the best version of the GTR model), and LG+C20+G4+F (again, the mixture model was used following the suggestion of Williams et al. (*28*); we were unable to use the higher-specified C60 model due to memory limitations of our hardware and used the C20 model instead).

**Ordination of asCOG phyletic patterns using Classical Multidimensional Scaling**

Binary asCOG presence-absence patterns were compared between pairs of Asgard MAGs using the following procedure: first, similarity between asCOG sets $\{\boldsymbol{A}\}$ and $\{\boldsymbol{B}\}$ was calculated as $S_{A,B} = |\boldsymbol{A} \cap \boldsymbol{B}|/\sqrt{|\boldsymbol{A}||\boldsymbol{B}|}$ (the number of shared AsCOGs normalized by the geometric mean of the number of AsCOGs in the two MAGs); then, the distance between the patterns was calculated as $d_{A,B} = -\ln(S_{A,B})$. The 162x162 distance matrix was embedded into a 2-dimensional space using Classical Multidimensional Scaling analysis implemented as the *cmdscale* function in R. The projection retained 89% of the original datapoint inertia.

**Identification and analysis of ESPs**

To identify eukaryotic signature proteins (ESPs), several strategies were employed. First, ESPs reported by Zaremba-Niedzwiedzka et al. (*29*) were mapped to asCOGs using PSI-BLAST (*15*). These asCOGs were additionally examined case by case using HHpred (*30*) with representative of the respective asCOG or the respective asCOG alignment used as the query. Second, all asCOGs were mapped to CDD profiles (*31*) using PSI-BLAST, hits to eukaryote-specific domains were selected. Most of the putative ESP asCOGs identified in this search, and all with E-value >1e-10, were additionally examined using HHpred, with a representative of the respective asCOG or the respective asCOG alignment as the query. Third, we analyzed frequently occurring asCOGs (present in at least 50% of Asgard genomes and in at least 30% of Heimdall genomes) that were not annotated automatically with the above two approaches using HHpred with a representative of the respective asCOG or the respective asCOG alignment as the query . Fourth, most of the putative ESP asCOGs detected with these approaches were used as queries for a PSI-BLAST search that was run for three iterations (with inclusion threshold E-value=0.0001) against an Asgard only protein sequence database. Additional unannotated asCOGs with similarity to the (putative) ESPs identified in this search were further examined using HHpred. Fifth, the genomic neighborhoods or all ESPs were examined, and proteins encoded by unannotated neighbor genes were analyzed using HHpred server.

**Metabolic pathway reconstruction**

The patterns of gene presence-absence in the asCOGs were used to reconstruct the metabolic pathways of Asgard archaea (Supplementary Table 8). The asCOGs were linked to the KEGG database and to the list of predicted metabolic enzymes of Asgard archaea reported by Spang et al. (*32*) (Supplementary Table 8). The classification of [NiFe] hydrogenases was performed by comparing the Asgard proteins of

cog.001539, cog.002254, cog.010021, cog.011939 and cog.012499 to HydDB (*33*). For phylogenetic analysis, the reference sequences of group 1, 3 and 4 [NiFe] hydrogenases were retrieved from HydDB. The sequences were filtered using cd-hit with a sequence identity cut-off of 90% prior to adding orthologous genes of cog.001539, cog.002254, cog.010021, cog.011939 and cog.012499 of Asgard archaea. All sequences for group 1, group 3 and group 4 [NiFe] hydrogenases were aligned using mafft-LINSI (*21*) and trimmed with BMGE (-m BLOSUM30 -h 0.6) (*22*). Maximum-likelihood phylogenetic analyses were performed using IQ-tree (*23*) with the best-fit model (group 1:LG + C60 + R + F, group 3: LG + C60 + R +F and group 4 LG + C50 + R + F), respectively, according to Bayesian information criterion (BIC). Support values were estimated using the SH-like approximate-likelihood ratio test and ultrafast bootstraps, respectively. We adapted a relaxed common denominator approach to determine the presence of certain pathway in one Asgard phyla (*32*), and combined with maximum parsimony principle (*34*) to infer the metabolisms of major ancestral forms.

## References

1. L. Cheng, T.-L. Qiu, X. Li, W.-D. Wang, Y. Deng, X.-B. Yin, H. Zhang, Isolation and characterization of Methanoculleus receptaculi sp. nov. from Shengli oil field, China. *FEMS Microbiology Letters*. **285**, 65–71 (2008).

2. J. Peng, Z. Lü, J. Rui, Y. Lu, Dynamics of the Methanogenic Archaeal Community during Plant Residue Decomposition in an Anoxic Rice Field Soil. *Appl. Environ. Microbiol.* **74**, 2894–2901 (2008).

3. G. V. Uritskiy, J. DiRuggiero, J. Taylor, MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome*. **6**, 158 (2018).

4. D. Li, C. M. Liu, R. Luo, K. Sadakane, T. W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*. **31**, 1674–1676 (2015).

5. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods*. **9**, 357–359 (2012).

6. D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, Z. Wang, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. **7**, e7359 (2019).

7. C. M. K. Sieber, A. J. Probst, A. Sharrar, B. C. Thomas, M. Hess, S. G. Tringe, J. F. Banfield, Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. **3**, 836–843 (2018).

8. Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin, IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. **28**, 1420–1428 (2012).

9. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**, 2114–2120 (2014).

10. S. Nurk, D. Meleshko, A. Korobeynikov, P. A. Pevzner, metaSPAdes: a new versatile metagenomic assembler. *Genome Res*. **27**, 824–834 (2017).

11. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*. **25**, 1043–1055 (2015).

12. D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **11**, 119 (2010).

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

14. P. P. Chan, T. M. Lowe, tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol Biol*. **1962**, 1–14 (2019).

15. A. A. Schäffer, L. Aravind, T. L. Madden, S. Shavirin, J. L. Spouge, Y. I. Wolf, E. V. Koonin, S. F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*. **29**, 2994–3005 (2001).

16. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol*. **35**, 1026–1028 (2017).

17. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. **5**, 113 (2004).

18. J. Söding, Protein homology detection by HMM-HMM comparison. *Bioinformatics*. **21**, 951–960 (2005).

19. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*. **5**, e9490 (2010).

20. P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*. **16**, 276–277 (2000).

21. K. Katoh, D. M. Standley, MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*. **30**, 772–780 (2013).

22. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evolutionary Biology*. **10**, 210 (2010).

23. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*. **32**, 268–274 (2015).

24. E. Esterman, Y. I. Wolf, R. Kogay, E. V. Koonin, O. Zhaxybayeva, *bioRxiv*, in press, doi:10.1101/2020.10.08.331884.

25. F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, P. Bork, Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*. **311**, 1283–1287 (2006).

26. P. Puigbò, Y. I. Wolf, E. V. Koonin, Search for a "Tree of Life" in the thicket of the phylogenetic forest. *Journal of biology*. **8**, 59 (2009).

27. A. Spang, J. H. Saw, S. L. Jørgensen, K. Zaremba-Niedzwiedzka, J. Martijn, A. E. Lind, R. van Eijk, C. Schleper, L. Guy, T. J. G. Ettema, Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature*. **521**, 173–179 (2015).

28. T. A. Williams, C. J. Cox, P. G. Foster, G. J. Szöllősi, T. M. Embley, Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol*. **4**, 138–147 (2020).

29. K. Zaremba-Niedzwiedzka, E. F. Caceres, J. H. Saw, D. B ckstr m, L. Juzokaite, E. Vancaester, K. W. Seitz, K. Anantharaman, P. Starnawski, K. U. Kjeldsen, M. B. Stott, T. Nunoura, J. F. Banfield, A.

Schramm, B. J. Baker, A. Spang, T. J. G. Ettema, Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*. **541**, 353–358 (2017).

30. L. Zimmermann, A. Stephens, S.-Z. Nam, D. Rau, J. Kübler, M. Lozajic, F. Gabler, J. Söding, A. N. Lupas, V. Alva, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol*. **430**, 2237–2243 (2018).

31. M. Yang, M. K. Derbyshire, R. A. Yamashita, A. Marchler-Bauer, NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. *Curr Protoc Bioinformatics*. **69**, e90 (2020).

32. A. Spang, C. W. Stairs, N. Dombrowski, L. Eme, J. Lombard, E. F. Caceres, C. Greening, B. J. Baker, T. J. G. Ettema, Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. *Nature Microbiology*. **4**, 1138–1148 (2019).

33. D. Søndergaard, C. N. S. Pedersen, C. Greening, HydDB: A web tool for hydrogenase classification and analysis. *Scientific Reports*. **6**, 34212 (2016).

34. D. L. Swofford, W. P. Maddison, Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*. **87**, 199–229 (1987).

## Data availability

Asgard archaea genomes generated in this study are available in eLMSG (an eLibrary of Microbial Systematics and Genomics, https://www.biosino.org/elmsg/index) under accession numbers from LMSG_G000000521.1 to LMSG_G000000609.1 and NCBI database under the project XXXXX.

Additional Data files are available at ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/asgard20/

Additional data file 1 (Additional_data_file_1.tgz): Complete asCOG data archive

Additional data file 2 (Additional_data_file_2.tgz): Phylogenetic trees and alignments archive

## Supplementary Figures

**Supplementary Figure 1** Global distribution of Asgard genomes analyzed in the study. The pie chart shows the proportion of Asgard genomes found in a biotope. Bold letters in the map show the sampling locations.

**Supplementary Figure 2** a. Distribution of completeness and contamination for 76 Asgard MAGs assessed by CheckM (v 1.0.12). Distribution of depth coverage (b) and N50 statistics (c) for Asgard MAGs reconstructed in this study. The numbers in parentheses indicate the number of Asgard genomes recovered from a given sampling location. The data for this plot can be found in **Supplementary Table 1.**

**Supplementary Figure 3** Comparison of the mean amino-acid identity (AAI) of Asgard and TACK superphyla. a. Shared AAI across Asgard and TACK lineages. Background: comparing all Asgard and TACK lineages included in the analyses but excluding archaea belonging to the same lineages and the six phyla proposed in the current work to investigate the distribution of AAI that defines a phylum. The AAI comparison of (b) Thorarchaeota, (c) Hermodarchaeota (d) Odinarchaeota, (e) Baldrarchaeota, (f) Lokiarchaeota, (g) Helarchaeota, (h) Borrarchaeota, (i) Heimdallarchaeota, (j) Kariarchaeota, (k) Gerdarchaeota, (l) Hodarchaeota and (m) Wukongarchaeota to other Asgad and TACK lineages. The lower and upper hinges of the boxplot correspond to the first and third quartiles. Data beyond the whiskers are shown as individual data points. Number in the parenthesis indicates the number of genomes in the lineages. Data for this plot is included in **Supplementary Table 2.**

**Supplementary Figure 4** Comparison of the 16S rRNA gene sequence (>1300 bp) identity of Asgard and TACK lineages. a. Shared 16S rRNA gene sequence identity across Asgard and TACK lineages. Background:comparing all Asgard and TACK lineages included in the analyses but excluding archaea belonging to the same lineages and the six phyla proposed in the current study to investigate the distribution of 16S rRNA gene identity that defines a phylum. 16S rRNA gene identity comparison of (b) Thorarchaeota, (c) Hermodarchaeota (d) Odinarchaeota, (e) Lokiarchaeota, (f) Helarchaeota, (g) Heimdallarchaeota, (h) Kariarchaeota, (i) Gerdarchaeota, (j) Hodarchaeota and (k) Wukongarchaeota to other Asgard and TACK lineages. The lower and upper hinges of the boxplot correspond to the first and third quartiles. Data beyond the whiskers are shown as individual data points. Number in the parenthesis indicates the number of 16S rRNA gene sequences compared in the lineages. Line represents a 16S rRNA gene identity of 75%. Data for this plot could be found in **Supplementary Table 3.**

**Supplementary Figure 5** Presence-absence of orthologs of Asgard core genes in other archaea, bacteria and eukaryotes.

**Supplementary Figure 6** Phylogenetic analysis of group 4 [NiFe] hydrogenases in the Asgard archaea. The unrooted maximum-likelihood phylogenetic tree was built from an alignment of 425 sequences including 110 sequences of Asgard archaea, with 308 amino-acid positions.

**Supplementary Figure 7** Phylogenetic analysis of group 3 [NiFe] hydrogenases in the Asgard archaea. The unrooted maximum-likelihood phylogenetic tree was built from an alignment of 813 sequences including 335 sequences of Asgard archaea, 514 amino-acid positions.

**Supplementary Figure 8** Phylogenetic analysis of group 1 [NiFe] hydrogenases in the Asgard archaea. The unrooted maximum-likelihood phylogenetic tree was built from an alignment of 541 sequences including 2 sequences of Wukongarchaeaota, with 745 amino-acid positions.

**Supplementary Figure 9** A schematic representation of the presence and absence of selected metabolic features in all (putative) phyla of Asgard archaea.

**Supplementary Figure 10** Gene structure of the contig encoding Group 1 [Ni,Fe]-hydrogenase in Wukongarchaeota. Abbreviations: Ftr, Formylmethanofuran:tetrahydromethanopterin formyltransferase; FwdD Formylmethanofuran dehydrogenase subunit D; FwdB, Formylmethanofuran dehydrogenase subunit B; FwdA, Formylmethanofuran dehydrogenase subunit A; FwdC, Formylmethanofuran dehydrogenase subunit C; HdrC, Heterodisulfide reductase, subunit C; HyaB, Ni,Fe-hydrogenase I large subunit; HyaA, Ni,Fe-hydrogenase I small subunit.

576 **Supplementary Tables**

577 **Supplementary Table 1** Genome information, proposed taxonomy and isolation data

578 **Supplementary Table 2** Mean amino-acid identity values (in %) comparing 66 TACK genomes and 184
579 Asgard genomes (162 high quality and 22 low-quality)

580 **Supplementary Table 3** The 16S rRNA gene sequence identity (in %) comparing TACK lineages and
581 Asgard lineages. The identity was calculated using sequences longer than 1300 bps

582 **Supplementary Table 4** Species and phyletic markers used for the tree of life reconstruction

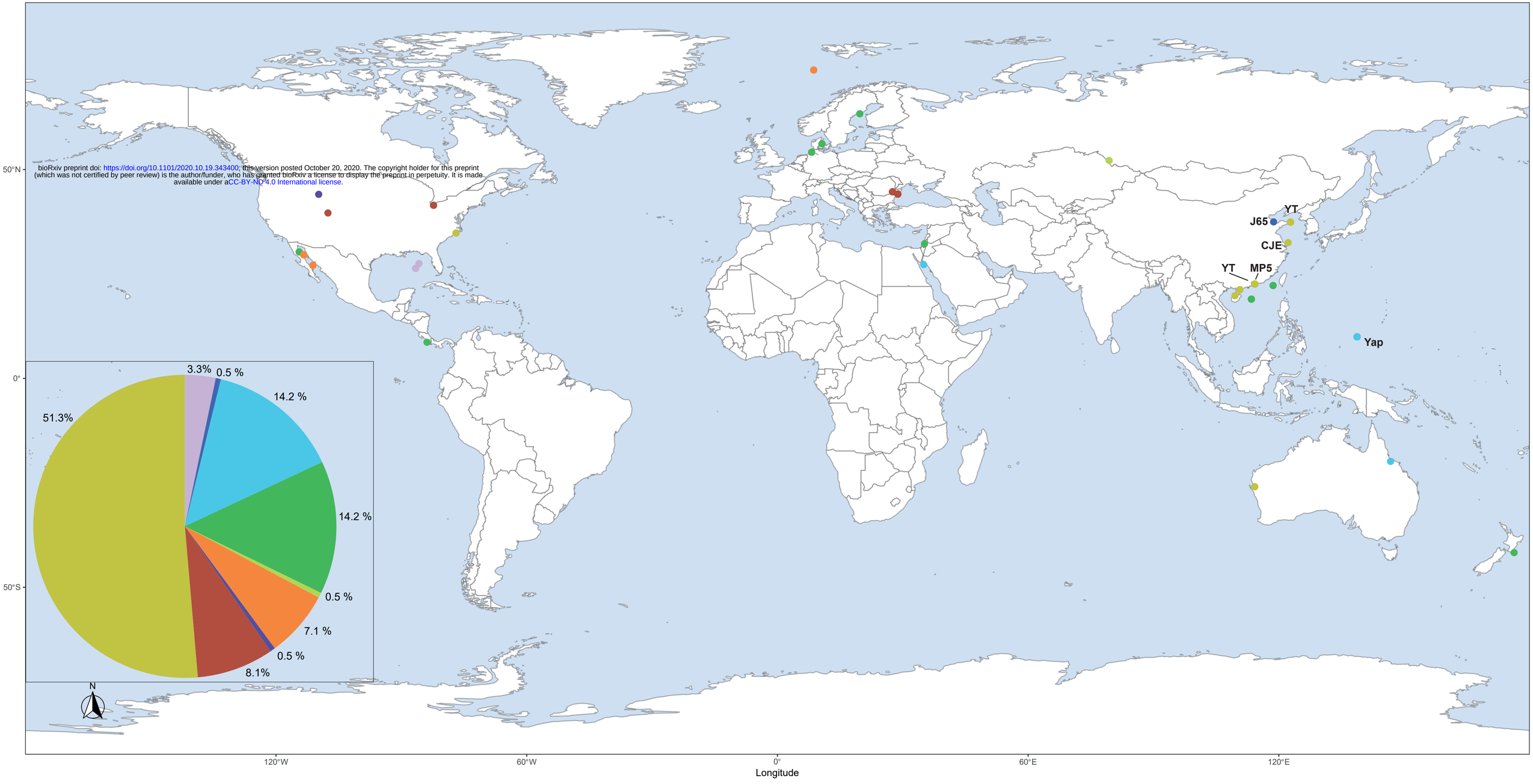
583 **Supplementary Table 5** Data for phylogenetic trees: the trees in the Newick format and the underlying
584 alignments

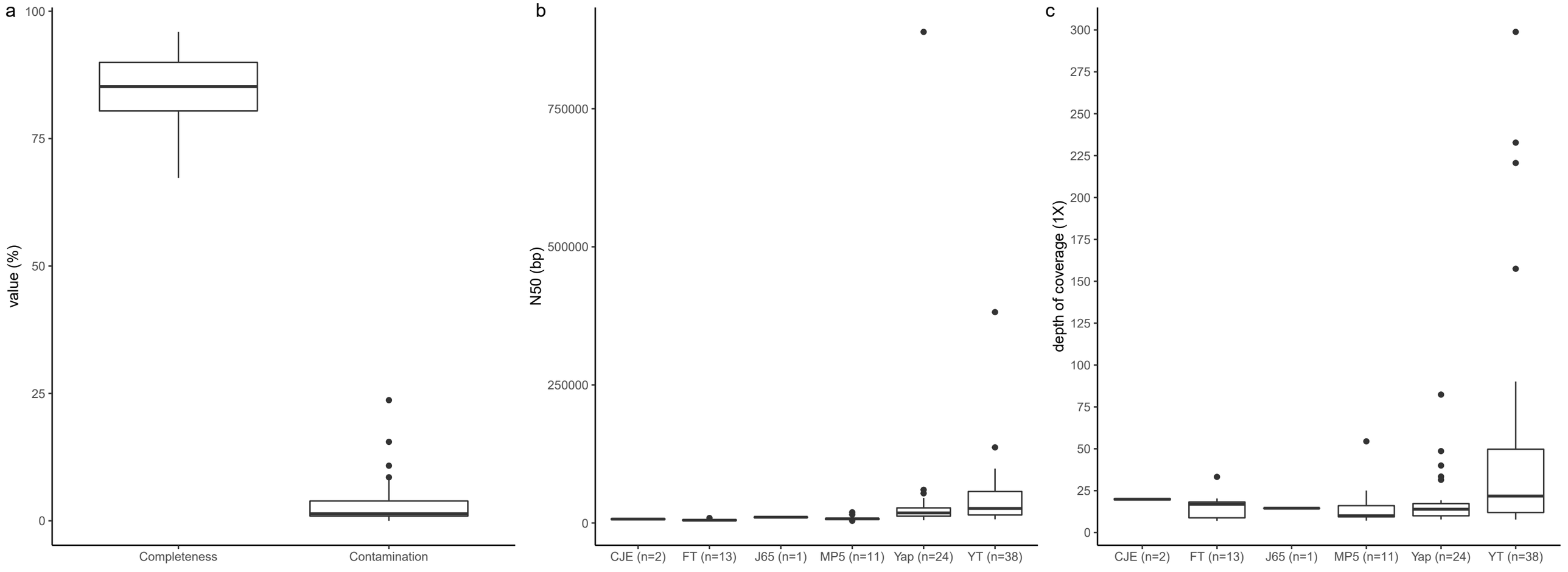
585 **Supplementary Table 6** The asCOGs annotation

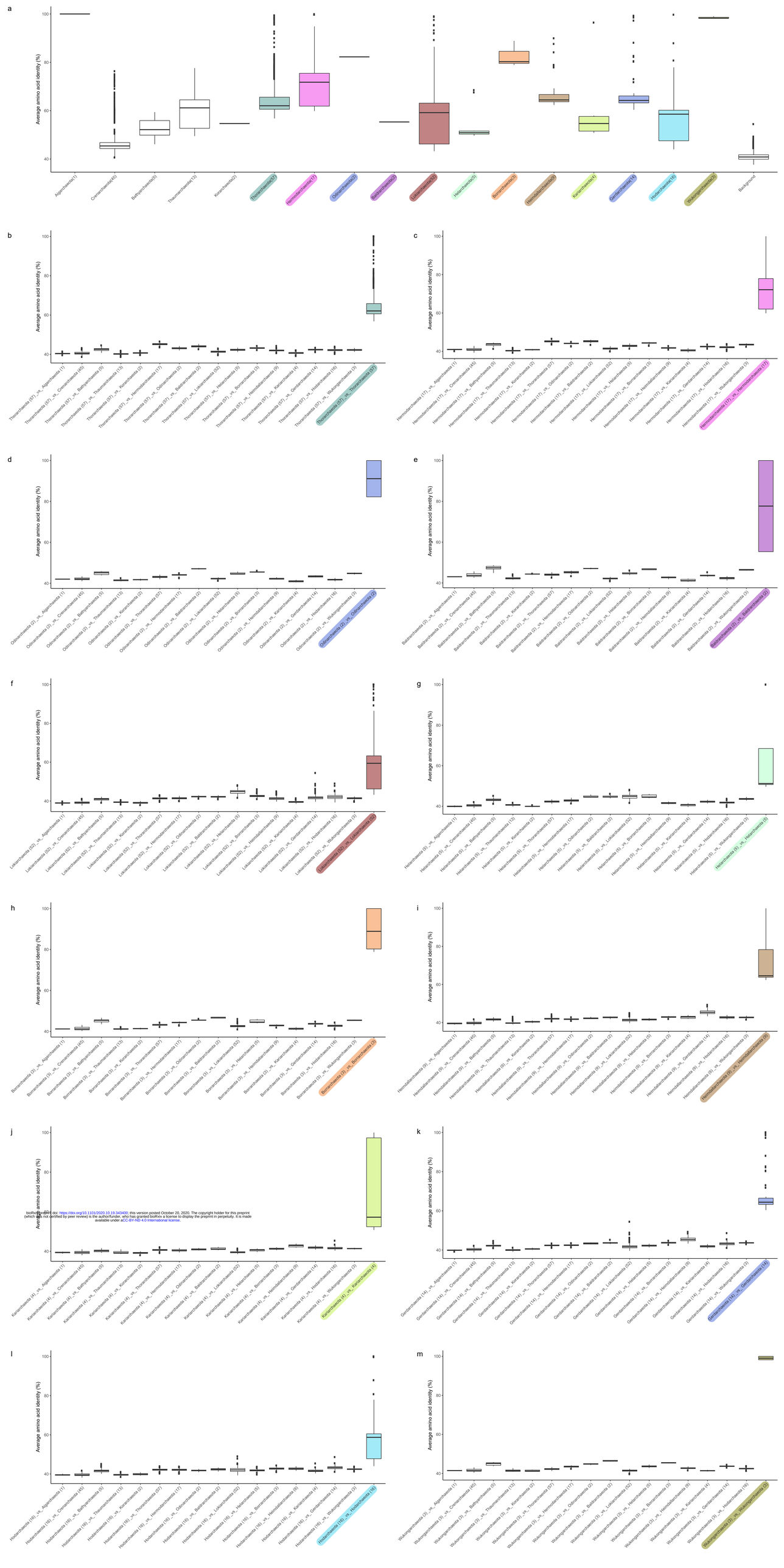
586 **Supplementary Table 7** Eukaryotic signature proteins in Asgard archaea

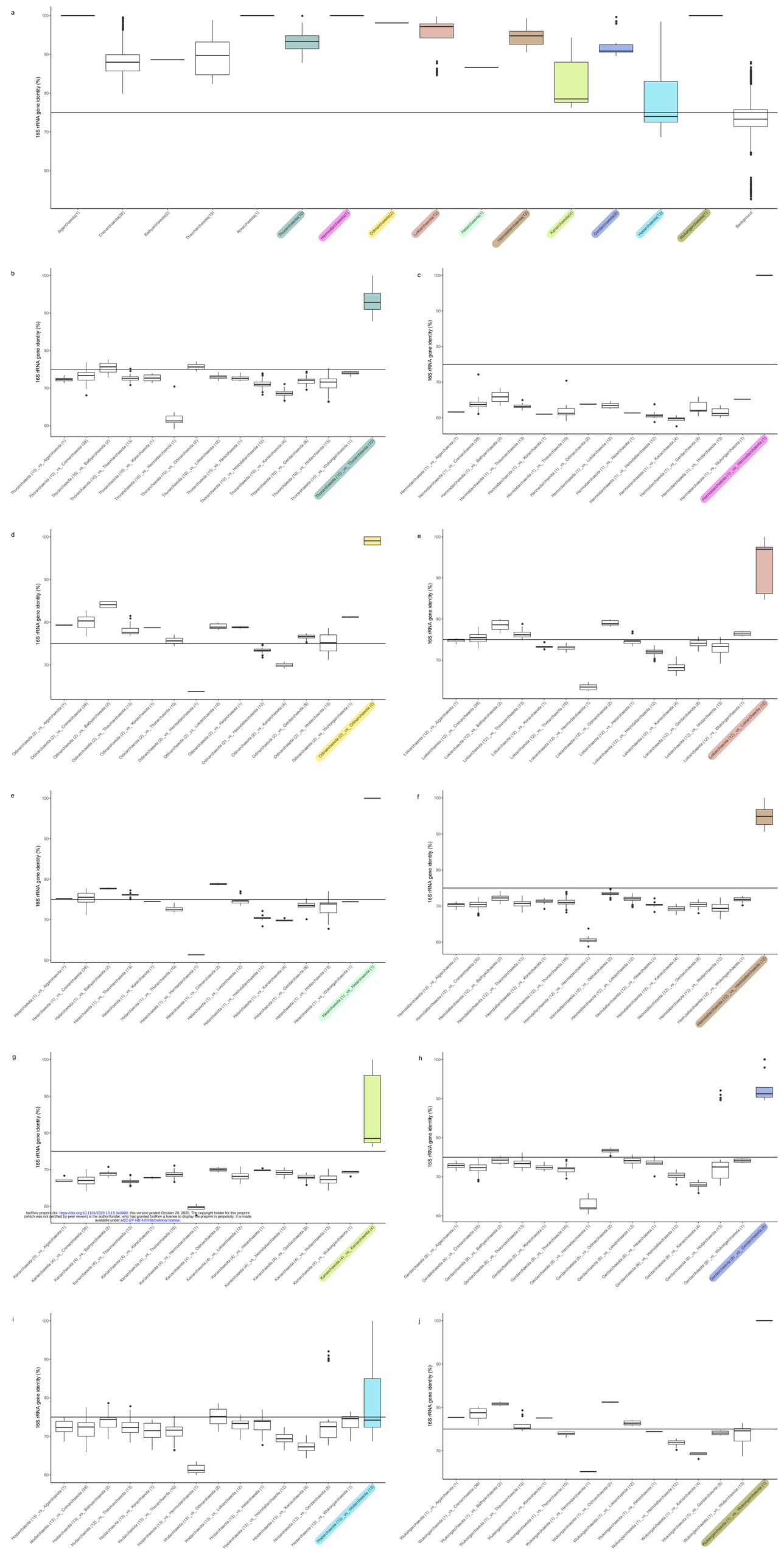
587 **Supplementary Table 8** The presence-absence of metabolic enzymes in Asgard archaea.

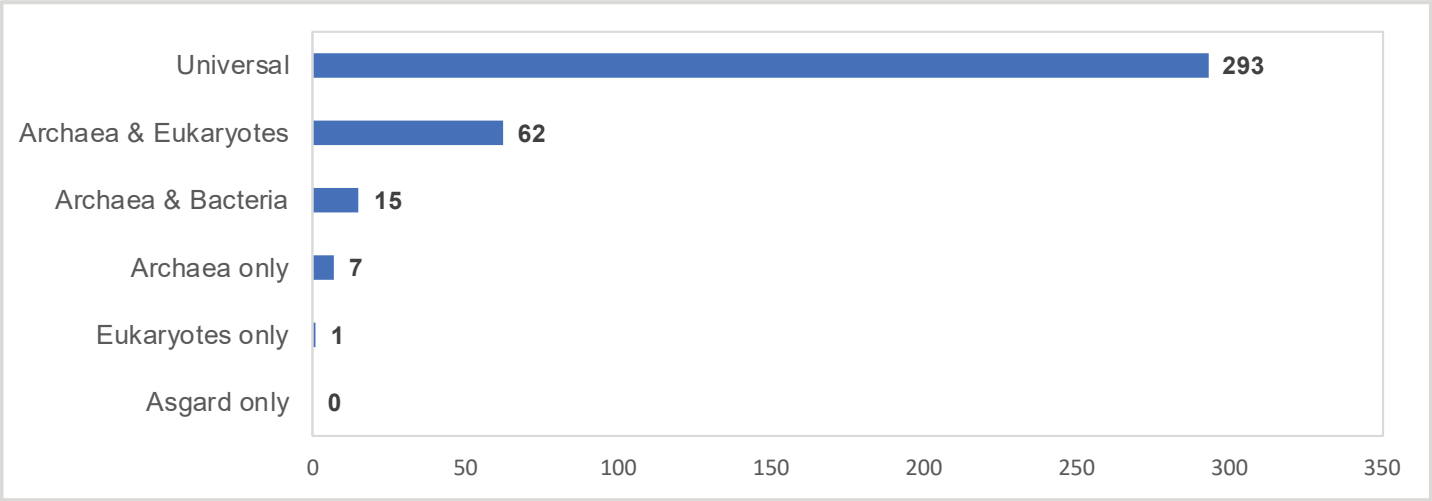
588

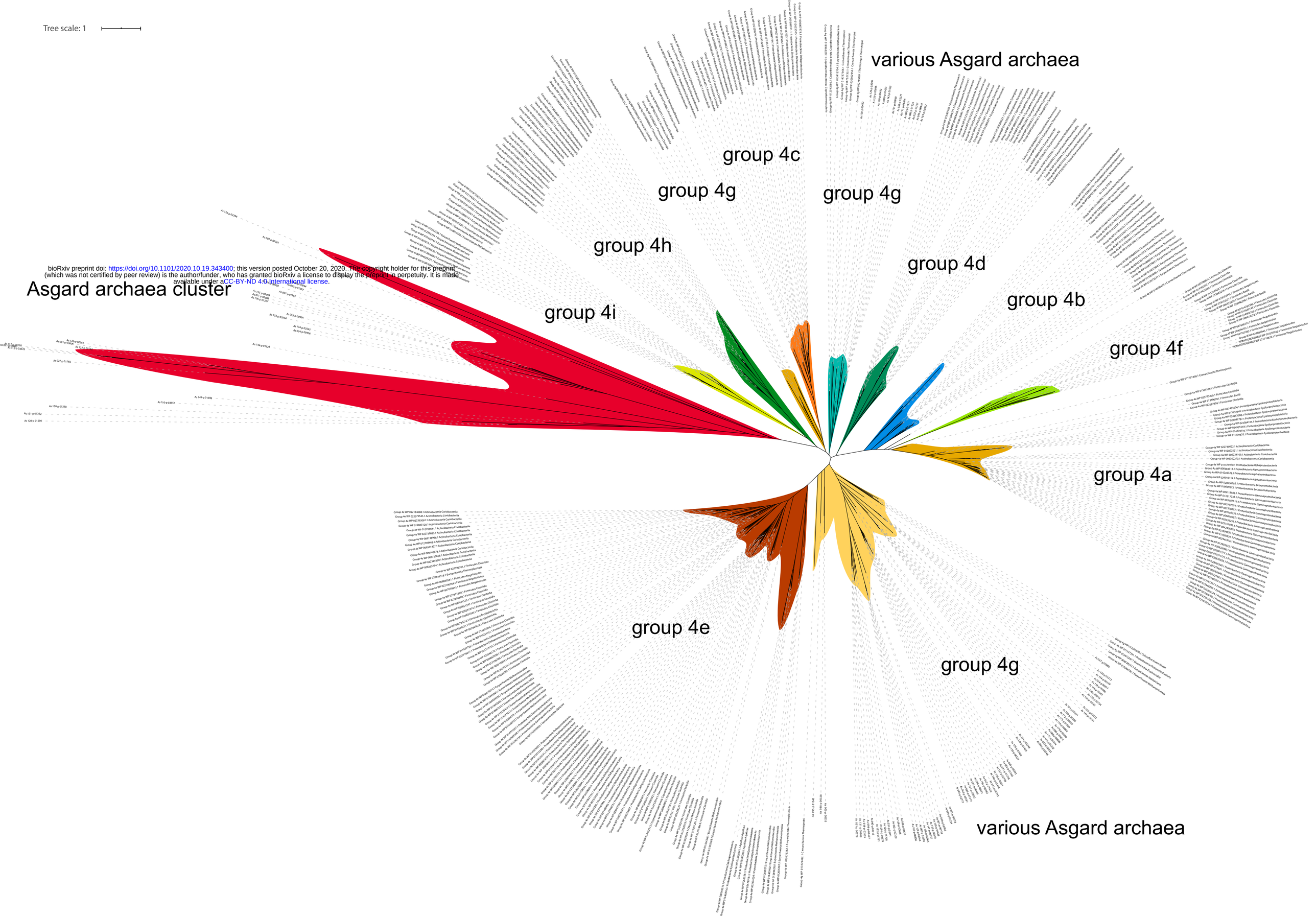


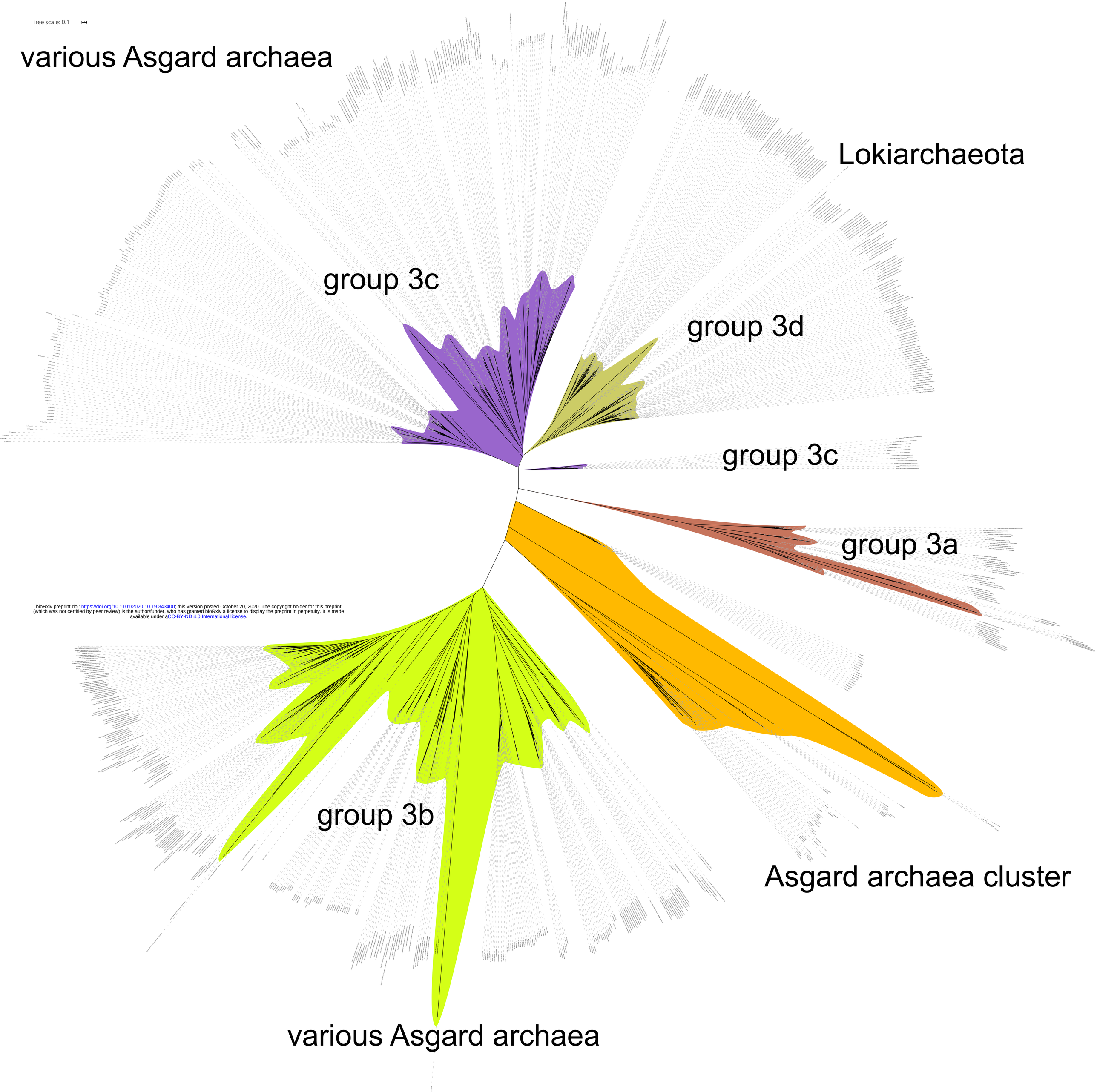

a
b
c
value (%)
N50 (bp)
depth of coverage (1X)
100
75
50
25
0
750000
500000
250000
300
275
250
225
200
175
150
125
100
75
50
25
0
Completeness
Contamination
CJE (n=2)
FT (n=13)
J65 (n=1)
MP5 (n=11)
Yap (n=24)
YT (n=38)

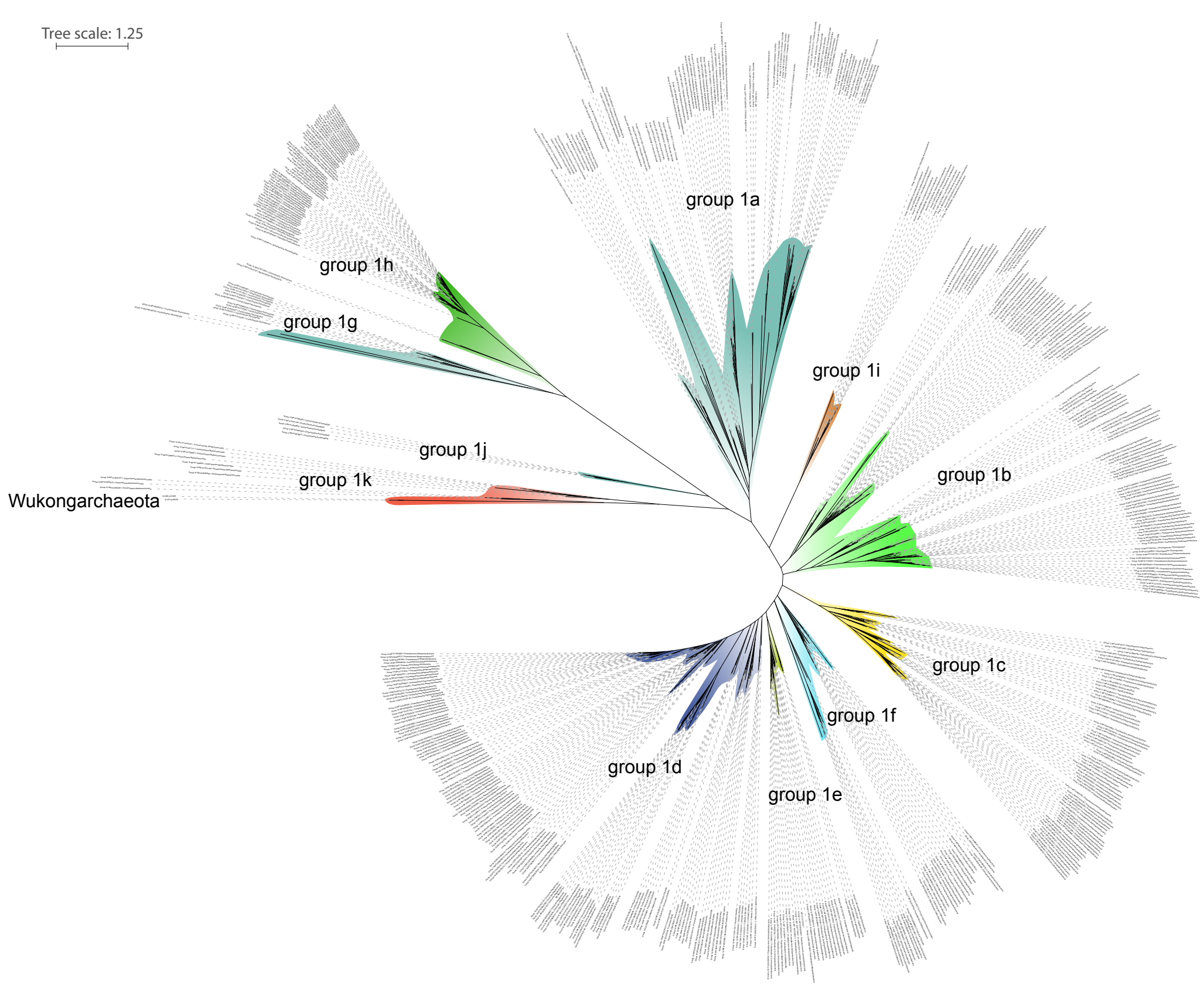


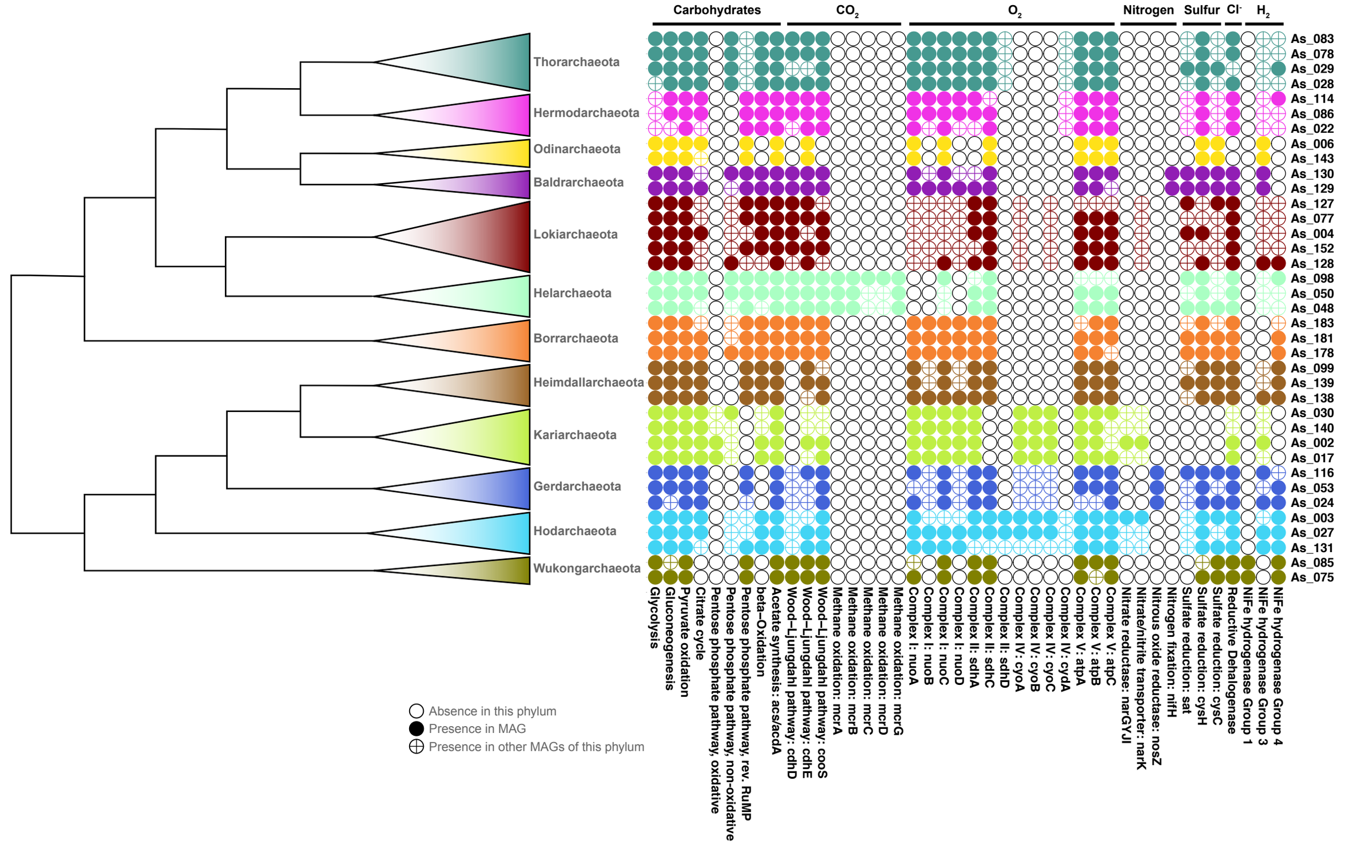


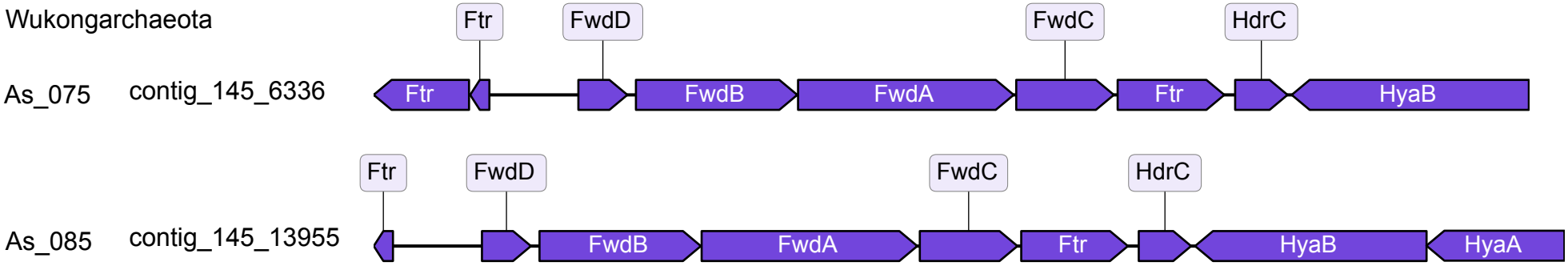







Tree scale: 1.25
group 1a
group 1h
group 1g
group 1i
group 1j
group 1k
group 1b
Wukongarchaeota
group 1c
group 1f
group 1d
group 1e


Carbohydrates
CO2
O2
Nitrogen
Sulfur
Cl-
H2
Thorarchaeota
Hermodarchaeota
Odinarchaeota
Baldrarchaeota
Lokiarchaeota
Helarchaeota
Borrarchaeota
Heimdallarchaeota
Kariarchaeota
Gerdarchaeota
Hodarchaeota
Wukongarchaeota
As_083
As_078
As_029
As_028
As_114
As_086
As_022
As_006
As_143
As_130
As_129
As_127
As_077
As_004
As_152
As_128
As_098
As_050
As_048
As_183
As_181
As_178
As_099
As_139
As_138
As_030
As_140
As_002
As_017
As_116
As_053
As_024
As_003
As_027
As_131
As_085
As_075
Glycolysis
Gluconeogenesis
Pyruvate oxidation
Citrate cycle
Pentose phosphate pathway, oxidative
Pentose phosphate pathway, non-oxidative
Pentose phosphate pathway, rev. RuMP
beta–Oxidation
Acetate synthesis: acs/acdA
Wood–Ljungdahl pathway: cdhD
Wood–Ljungdahl pathway: cdhE
Wood–Ljungdahl pathway: cooS
Methane oxidation: mcrA
Methane oxidation: mcrB
Methane oxidation: mcrC
Methane oxidation: mcrD
Methane oxidation: mcrG
Complex I: nuoA
Complex I: nuoB
Complex I: nuoC
Complex I: nuoD
Complex II: sdhA
Complex II: sdhC
Complex II: sdhD
Complex IV: cyoA
Complex IV: cyoB
Complex IV: cyoC
Complex IV: cydA
Complex V: atpA
Complex V: atpB
Complex V: atpC
Nitrate reductase: narGYJI
Nitrate/nitrite transporter: narK
Nitrous oxide reductase: nosZ
Nitrogen fixation: nifH
Sulfate reduction: sat
Sulfate reduction: cysH
Sulfate reduction: cysC
Reductive Dehalogenase
NiFe hydrogenase Group 1
NiFe hydrogenase Group 3
NiFe hydrogenase Group 4
Absence in this phylum
Presence in MAG
Presence in other MAGs of this phylum


Wukongarchaeota
Ftr
FwdD
FwdC
HdrC
As_075
contig_145_6336
Ftr
FwdB
FwdA
Ftr
HyaB
Ftr
FwdD
FwdC
HdrC
As_085
contig_145_13955
FwdB
FwdA
Ftr
HyaB
HyaA