Phylogenomics of Asgard archaea reveals a unique blend of prokaryotic-like horizontal transfer and eukaryotic-like gene duplication.

Saioa Manzano-Morales[1,2], and Toni Gabaldón[1,2,3.4*]

1) Barcelona Supercomputing Center (BSC). Plaça Eusebi Güell, 1-3

08034 Barcelona, Spain.

2) Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, 08028 Barcelona, Spain

3) Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain

4) CIBER de Enfermedades Infecciosas, Instituto de Salud Carlos III, Madrid, Spain

*Corresponding author: toni.gabaldon@bsc.es

Abstract

Asgard archaea hold a pivotal position in the tree of life as the closest known relatives to eukaryotes and are therefore crucial for understanding eukaryogenesis. Earlier genomic analyses revealed that Asgard genomes are remarkably larger than those of other archaea and contain a significant number of genes seemingly acquired from bacteria. However, the precise contributions of horizontal gene transfer and gene duplication in shaping Asgard genomes remain largely unknown. Here, we present a comprehensive phylogenomic analysis to dissect the evolutionary dynamics of Asgard genomes, quantifying gene duplication, loss, and both inter- and intra-domain gene transfer events. Our findings reveal that gene transfer is widespread throughout Asgard evolution, predominantly affecting metabolic genes at the periphery of interaction networks. However, our analyses demonstrate that gene duplications, rather than horizontal gene transfers, are the primary drivers behind the increased genome sizes observed in Asgard archaea. This unique evolutionary signature in Asgard archaea—a blend of pervasive prokaryotic-like gene transfer alongside significant eukaryotic-like gene duplication—is consistent with their phylogenetic placement and offers novel insights into the genomic transitions that likely underpinned eukaryogenesis.

Introduction

Asgard archaea are the closest prokaryotic relatives of eukaryotes and are therefore key to understanding eukaryogenesis[1,2]. Asgard genomes encode many of what were previously thought to be eukaryote-specific proteins[2], including actin and actin-related proteins[3], and an actin cytoskeleton has been found in a cultivated Lokiarchaeum[4]. The crucial interest of Asgard archaea has promoted sequencing efforts, resulting in a large number of available genomes - 229 assemblies in release 207 of Genome Taxonomy Database (GTDB)[5]. These represent 10 class-level lineages *sensu* GTDB (namely Baldr-, Heimdall-, Hermo-, Jord-, Loki-, Odin-, Sif-, Thor-, and Wukongarchaeia, as well as a LC30 clade). However, most genomic data corresponds to Metagenomic Assembled Genomes (MAGs), which can be contentious due to potential incompleteness or contamination issues[6] . Asgard archaea have proven elusive to culture, with two exceptions from Lokiarchaea: *Ca.* Prometheoarchaeum syntrophicum MK-D1[7] and *Ca.* Lokiarchaeum ossiferum B-35 4, which can be cultured with their syntrophic partners and therefore have enabled high-quality genomes.

Asgard archaea were first identified in a deep-sea hydrothermal vent[8], but their presence in clone libraries precedes their description by two decades[9–11]. Recent sampling efforts have uncovered new Asgard lineages, of which most were first described in anoxic marine environments[12]. However, metagenomic analyses indicate a worldwide cosmopolitan distribution[13]. This reflects on a broad array of lifestyles and metabolic properties including mixotrophy, homoacetogenesis, facultative anaerobiosis, alkane utilization, and rhodopsin-based phototrophy[13–17]. However, we have a very limited understanding of how genome variation underlies such diverse adaptations.

Earlier analyses of Asgard genomes uncovered a large number of putatively transferred bacterial genes[18] and generally larger genome sizes as compared to other archaea[19]. Previous analyses considering prokaryotes have found weak[20] to strong[21,22] positive correlations between genome size and horizontal gene transfer (HGT) fraction, but none of them included Asgard genomes. Similarly, Asgard archaeal genomes have been shown to have higher rates of duplication than other archaea[19], but the functional implications of these gene duplications, or how these duplication tendencies contrast with HGT (the prevailing form of genome expansion in prokaryotes) have not been comprehensively explored .

Here, we reconstructed the Asgard archaeal pangenome and used phylogenomics on a super-phylum-wide scale to assess the relative contribution of gene duplication, loss and

transfer in the evolution of Asgard genomes, and investigate possible functional trends. Our results indicate that Asgard archaea have high duplication rates that are intermediate between those in their prokaryotic sister clade (despite similar transfer rates) and those observed in unicellular eukaryotes. We uncover widespread but mostly lineage-specific inter-domain HGT involving several donor-acceptor pairs and metabolic genes. Overall, our results indicate that Asgard genomes are uniquely shaped by a blend of prokaryotic horizontal gene transfer and eukaryotic-style gene duplications.

Results

The pangenome and signature proteins of Asgard archaea

We first sought to characterize the Asgard pangenome by performing homology-based clustering of all available genomes from phylum Asgardarchaeota in GTDB[5]. However, nearly all Asgard genomes are MAGs, resulting in a median CheckM[23] completeness of 87.31%. This lack of genome completeness can lead to an infra-estimation of the core genome and, conversely, an over-estimation of the accessory genome.

To circumvent this, we used a representative subset of 31 genomes that were chosen among the most complete genomes from each one of the major Asgard lineages, namely Baldr-, Heimdall-, Hermod-, Hel-, Jord-, Loki-, Odin-, Sif-, Thor-, and Wukongarchaeia, as well as Hodarchaeales due to its importance in relation to eukaryogenesis (See **Online Methods, Extended Data Table 1**). These lineages are all class-level *sensu* GTDB, except for Hod- and Helarchaeales, and will be referred to as "lineages" throughout the text. Additionally, we included 30 other prokaryotic genomes as out-groups, including 17 representative species from the Asgard archaea's sister group, TACK (originally standing for Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota, now comprising more clades and named Thermoproteota *sensu* GTDB).

We clustered proteins into families (orthogroups, OGs) using a similarity-based approach (See **Online Methods**). This yielded a total of 31,083 Asgard archaeal OGs, including singletons, out of which 17,234 contained at least one Asgard representative, 13,636 were exclusively present in Asgard archaea, and 8,365 were present in a single Asgard genome and no other prokaryote in our dataset. 188 gene families are found in all Asgard genomes (strict core) and 473 in 90% or more (relaxed core), with the remaining belonging to the accessory pangenome. These relaxed core proteins likely carry out key processes for the functioning of

the cell, and we found they were enriched in KEGG pathways such as ribosome (ko03010), DNA replication (ko03030), carbon metabolism (ko01200) and RNA polymerase (ko03020) (**Extended Data Table 2**).

We estimated the size of the total Asgard pangenome both by Chao's lower bound[24] (yielding a lower bound estimate of 31,239 OGs) and by the best-fitting binomial mixture model (resulting in a pangenome of 56,323 OGs and a strict core of 111 OGs). In this binomial mixture model, the pangenome is categorized into seven bins, with the majority (99.8%) of the pangenome corresponding to accessory OGs (Fig. **1A,B**) and a comparatively small strict core (0.198%). We additionally calculated rarefaction curves (Fig. **1C**) which, along with the Heaps estimate (alpha = 0.34), indicated an open pangenome. To assess the extent to which partial genome completeness in MAGs impacts Asgard pan-genome size estimates, we performed the same analyses but now replacing the 31 best-quality representatives by 31 taxonomically-equivalent genomes chosen randomly from the entire genome set. In this case, the number of OGs is 32,692, of which 18,972 include at least one Asgard representative, and yields a similarly open pangenome with a Heaps alpha of 0.38, but a higher Chao's lower bound (46,672) and smaller total estimated size (50,439), and strict core (54). Hence, both genomic subsets provided similar estimates in terms of openness, OGs and total pan-genome size but not in terms of core genome or lower bound estimates, which were more affected by genome incompleteness.

We next used the full set of species representatives and the same outgroups to assess pan-genome estimates. In this case we obtained 61,882 OGs with at least one Asgard representative, Heaps alpha 0.36, a lower bound of 166,502 gene families and an estimated size of 131,957 gene families. These results, even considering the most conservative estimates, underscore our very partial knowledge of the breadth of Asgard archaeal protein diversity, implying the intriguing possibility that new protein families with potential implications for eukaryogenesis may await discovery.

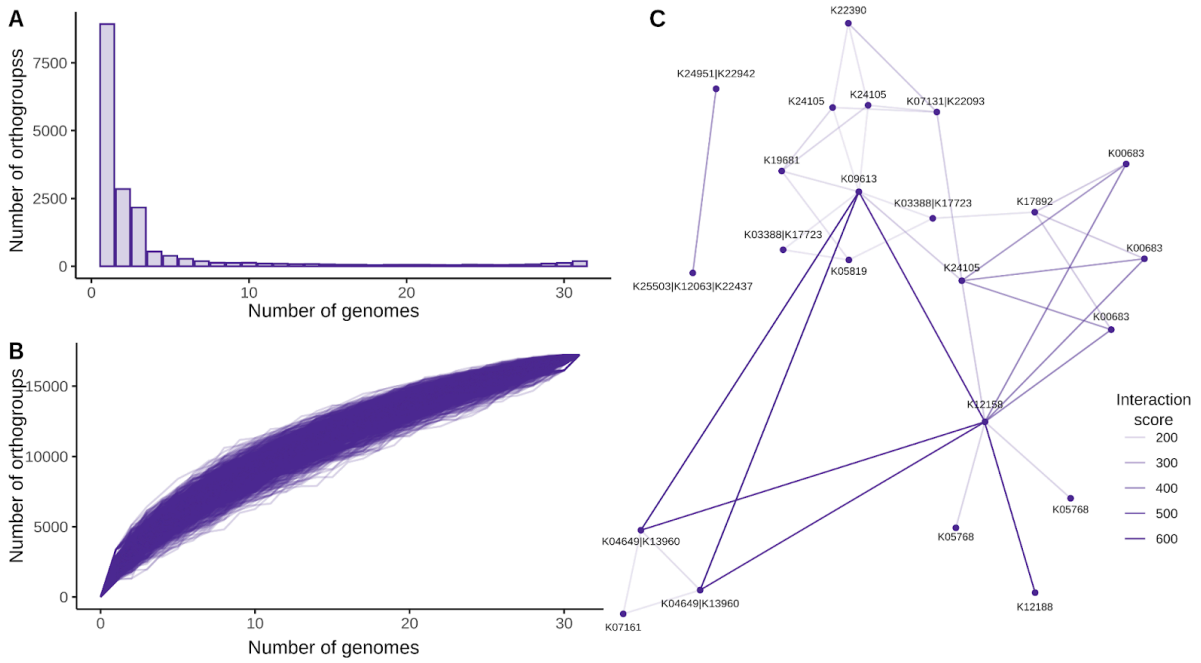


Figure 1: Asgard archaeal pangenome metrics for the high-quality subset. (A) Frequency of OGs according to the number of genomes in which they are present in the pangenome of Asgard archaea (bin width=1). (B) Rarefaction curves obtained by sampling the entire genome set in a random order over 1000 iterations . (C) Protein-protein interaction network of the Asgard-Signature proteins. Each node is one protein (labeled as its annotated KO/KOs), edges represent protein-protein interactions with color strength proportional to the interaction score given by STRING.

We defined Asgard signature proteins (ASPs) as those OGs common (present in ≥ 70% of the proteomes) in Asgard archaea but rare (present in ≤15% of the proteomes) in other archaea or bacteria. This rendered 53 ASPs, corresponding to 72 proteins in the Hermodarchaeia representative (the one with the highest number of ESPs in its proteome, whose PPI network we used as a template), which were found to be highly connected in protein-protein interactions networks **(Figure 1D)** and enriched in endocytosis (ko04144, p.adjust 5.284261e-06) and Motor protein (ko04814, p.adjust 8.09e-03) functions (**Extended Data Table 2**).

We cross-mapped a set of 83 eukaryotic Signature proteins (ESPs) recently identified in Asgard archaea[19] and found 39 of them in our dataset. Eleven such ESPs correspond to ASPs, suggesting that these eukaryotic-like proteins contributed, together with other ASPs, to the

differentiation of Asgard archaea from other prokaryotes and that some of these differentiating gene families ended up into the eukaryotic lineage.

To identify lineage-level signature proteins that may underlie lineage-specific adaptations, we selected OG that are widespread within each lineage (≥ 70% of genomes) and rare outside (≤15%). These gene families were enriched in "two-component system" (ko02020) in Thorarchaeia and Wukongarchaeia, and "cell cycle" and "quorum sensing" (ko04110, ko02024) in Lokiarchaeia signature proteins (**Extended Data Table 2**).

Genome dynamics across Asgard archaea

To assess dynamic changes and their underlying mechanisms in Asgard genomes, we performed a comprehensive phylogenomic analysis. For this we selected twelve representative Asgard species and used them as seeds for the reconstruction of the evolutionary histories of all their genes in the context of the 61 prokaryotic genome dataset described above (31 Asgards, 31 non-Asgard prokaryotes, including 17 TACK). These 12 phylomes comprised a total of 32,643 gene phylogenies. We additionally ran the same pipeline over a set of 8 seed TACK genomes against the same prokaryotic genome dataset.

We then analyzed the resulting gene trees with both species-overlap[25] and reconciliation[26] algorithms using two alternative species tree topologies for the Asgard clade (see **Online Methods**) to i) detect and date gene duplication events, including gene family expansions ii) infer orthology and paralogy relationships, and iii) infer HGT.

The average number of duplications per gene and lineage (duplication rate hereafter), as computed with either the species-overlap or reconciliation algorithms were significantly correlated (p-value 0.0014, R 0.49), and we show reconciliation-based duplication rates, together with transfer and gene loss rates in **Fig. 2A**. Our results confirmed previously described high levels of duplications at the base of Heimdallarchaeia and Lokiarchaeia[19], but extended this observation throughout the Asgard phylogeny. Duplication rates were particularly high at the base of the Asgard group and at the branch separating the majority of Asgard lineages from the earliest-diverging clades (Baldrarchaeia and Odinarchaeia), with additional hotspots at the base of Helarchaeales and the above-mentioned Lokiarchaeia, as well as Heimdallarchaeia, a key lineage to eukaryogenesis as current reconstructions suggest eukaryotes either stemmed from within[19] or as sister to[27] this clade. Some of the bursts of duplications coincide with the origin of clades with larger genomes, hinting towards gene

duplication driving increased genome sizes in Asgard archaea. Conversely, lineages with smaller genome sizes are associated with high rates of gene loss. Importantly, the high duplication rates in Asgards were the result of large expansions in a few gene families (2% of the genes have duplicated at least once) **(Fig. 2B)**, which were functionally diverse with only one functional term ("ribosome", ko03010) being enriched. Compared with TACK, Asgards display significantly higher levels of gene duplication despite having no significant differences in HGT levels **(Fig 2C,D),** suggesting that they might display duplication levels more akin to those in eukaryotes, yet maintain prokaryote-level HGT. To test this, we calculated the gene phylogenies of all the genes in a set of 4 eukaryotic seed genomes against a dataset of 30 Free-Living Unicellular Heterotrophic eukaryotes (FLUHs) and calculated duplication rates for FLUHs, Asgards and TACK with the species-overlap algorithm **(Fig. 2E)**. Asgard archaea displayed duplication rates that were intermediate between those in TACK and free-living unicellular eukaryotes, with the difference between Asgards and FLUHs being non-significant statistically.

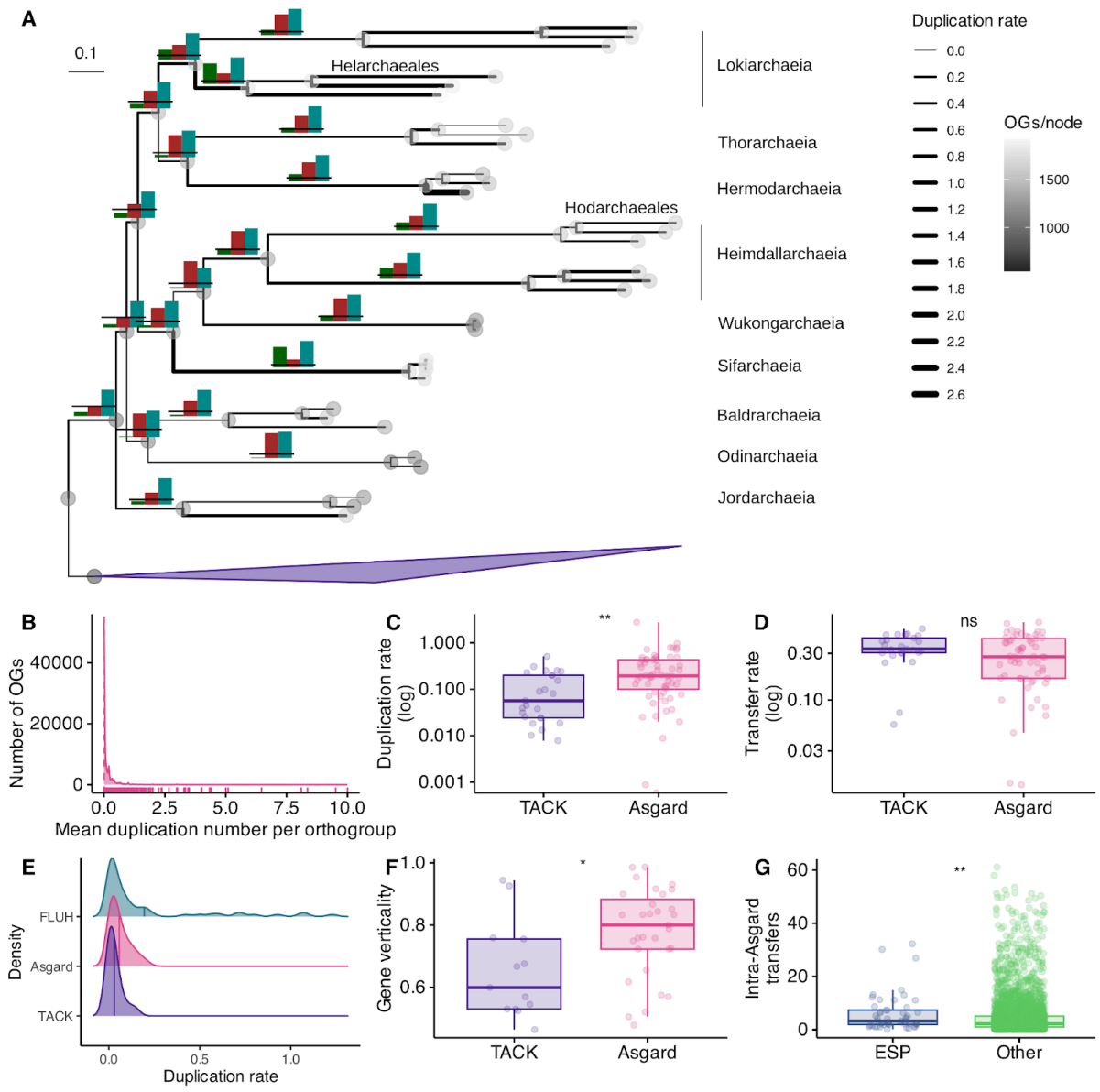


Figure 2: Gene gain-loss dynamics across Asgard archaea. (A) Gain and loss dynamics in Asgard archaea. The total number of OGs per proteome is displayed by circle color in each node. Duplication rate is represented by the branch width. Number of duplications (green), transfers (blue) and losses (red) displayed at internal nodes in absolute number. Horizontal bar indicates 100 events as scale. (B) Mean number of duplication events per node. A value of 1 would indicate a whole genome duplication. Duplication (C) and transfer (D) rates per node of Asgard and TACK gene families. (E) Duplication rates, inferred by species overlap, of Asgard, TACK and FLUHs. Black vertical lines mark 100 events, for each scale. (F) Gene verticality in TACK and Asgard archaea. (F) Intra-Asgard transfer events on ESPs compared to the rest of gene families.

Next, we looked at the set of "lineage signature proteins" described above to understand the relative contribution of gene duplication to lineage diversification. We compared the duplication rates of these OGs within and outside the lineage and we found that, for four lineages including the ones with highest overall duplication rates, these signature OGs were significantly more duplicated within than outside the lineage (**Extended Data Figure 1**), hinting to a possible role of gene duplication in lineage-specific adaptations in these lineages.

We searched for tandem duplications (e.g. resulting in sets of paralogous genes that are contiguous in the genome), and found a highly variable number across Asgard genomes (**Extended Data Figure 2A-C**). Closer inspection of tandem duplications, revealed interesting cases of the same gene, or block of genes, being repeatedly duplicated in tandem , as well as cases of different tandem duplications occurring contiguously in the genome **(Extended Data Figure 2D)**. Genes duplicated in tandem were enriched in quorum sensing (ko02024) in genomes of Heimdallarchaeia and Lokiarchaeia (GCA_014730275.1, GCA_013166775.1, GCA_025839675.1, GCF_008000775.1), two-component systems (ko02020) in Heimdallarchaeial genome GCA_014730275.1 and ABC transporters (ko02024) in Heimdallarchaeial and Lokiarchaeial genomes (GCA_014730275.1, GCA_025839675.1, GCF_008000775.1).

Horizontal Gene Transfer in Asgard archaea

We then assessed HGT, both intra- and inter-domain, across Asgard archaeal genomes. We first analyzed HGT using gene tree reconciliation algorithms in the above described phylogenetic dataset (see **Online Methods**). We found that, overall, Asgard archaea gene phylogenies showed a higher average verticality (0.76) than its sister clade TACK (0.63) **(Figure 2F**), mostly due to higher duplication rates despite comparable levels of transfer (**Fig 2C,D)**. Among Asgard archaea, all lineages had similar levels of HGT (Kruskal-Wallis test, p=0.29), while the duplication rates were lineage-dependent (Kruskal-Wallis p=0.029).

We focused on intra-Asgard HGT exchange partners. HGT patterns depict a dense network, in which most partners have exchanged a comparatively low number of genes, but where certain key pairs stand out (**Figure 2A and Extended Data Figure 3**), such as highways between distant clades like Jord and Helarchaeales. For the top 5 transfer pairs, we ran a KEGG pathway enrichment of the orthogroups transferred with probability ≥ 0.7 and obtained an enrichment in metabolic functions. Overall, we interpret these results as evidence of a high degree of phylogenetic incongruence in the context of a strong phylogenetic

backbone, with frequent and dispersed HGT. Importantly, intra-Asgard HGT has been proposed as a potential factor explaining the absence of ESPs in Heimdallarchaeia, the closest Asgard relatives of Eukaryotes[28]. In this regard, we did detect a significantly higher rate of involvement in intra-Asgard HGT of ESPs as compared to other genes (**Fig. 2G**).

To assess inter-domain HGT we performed an exhaustive search restricted to the twelve high-quality seed Asgard genomes, by using a combined BLAST- and phylogeny-based approach. as well as using a broader database comprising 491,850 sequenced genomes from prokaryotes, eukaryotes and viruses (see **Online Methods**).

Blast-based approaches suggested a large amount of HGT, involving between 12-21% of the total gene repertoire (**Extended Data Table 3**). As many blast-based HGT predictions are likely to be false positives, we reconstructed phylogenies for each candidate and ran two different phylogeny-based HGT detection algorithms (see **Online Methods**). This approach predicted 3,628 gene transfer events in the 12 investigated genomes, involving 2-14% of the gene repertoire. The overall good agreement between the two phylogeny-based HGT detection approaches, the generally high phylogenetic support (**Extended Data Figure 4A**), the confirmation of a previously described HGT event[29], and the lack of correlation with genomic contamination estimates (**Extended Data Figure 4B**), suggested that our pipeline is rather conservative and that our phylogeny-based inter-domain HGT inferences are likely to be correct. We additionally kept a subset of transfer events for which the bootstrap was well-supported (≥70), which we labeled as the “high-confidence” set, to test that the qualitative conclusions held at higher stringency.

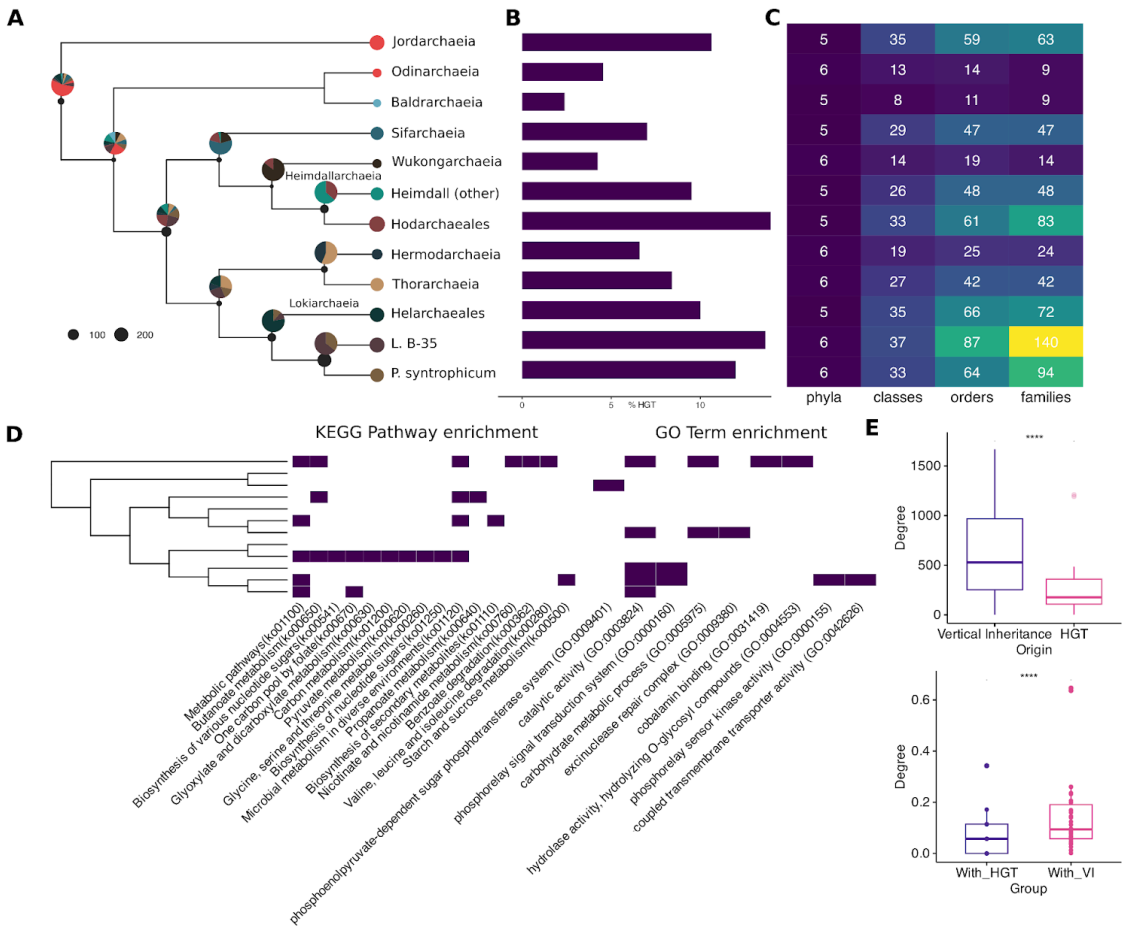


Figure 3: HGT-derived proteins across the Asgard lineage. A) Prevalence of HGT across the Asgard phylogeny. Dot size represents the mean number of gene trees with HGT events mapped as having occurred in that given ancestor, average of the HGT events detected per seed. Pie chart above dot represents relative contribution of the different seeds to the inferred HGT events. B) Percentage of genes involved in HGT, relative to the total number of protein-coding genes per lineage. C) Number of donor phyla per lineage that have contributed at least 3% of the HGT genes to the given lineage, and their taxonomic composition at class, order and family level (D) KEGG KO and GO term enrichment for the proteins derived from Bacteria-to-Asgard HGT. (D) Network properties of HGT-acquired genes versus vertically inherited ones: degree (number of interactions) of the proteins of HGT origin and of VI origin (above); for the HGT proteins, we show the number of connections per node of the subnetwork composed only of HGT proteins (with_HGT) and calculate the number of connections with VI proteins

Based on the gene tree topologies, we inferred the directionality of the transfers (**Extended Data Figure 4C**), used the last common ancestor of the genomes carrying the transferred genes as a proxy for the relative time of the acquisition, and mapped this onto the Asgard species tree (**Fig. 3A**). Although our pipeline was designed to primarily search for Bacteria-to-Asgard gene transfers, we captured some Asgard-to-Bacteria transfers (**Extended Data Figure 3C**).

We assessed the relative contribution of the different bacterial clades as exchange partners, Our results (**Figure 3C**) show many different partners, with varying levels of taxonomic width. Interestingly, some of the identified donors belong to taxonomic groups that are known to co-occur or interact with extant Asgards. For example, Lokiarchaeota co-occurs with Deltaproteobacteria (phylum Proteobacteria) and Anaerolineaeae (phylum Chloroflexota), both of which appear as significant transfer partners (>5% of trees). Desulfobacteriota is particularly interesting in the Lokiarchaeia lineages, as both *Ca.* P. syntrophicum and *Ca.* L. ossiferum have known syntrophic partnerships with sulfate-reducing bacteria, of the *Desulfovibrio* and *Halodesulfovibrio* genera[4,7]. Importantly, most of the Desulfobacterota HGT partners identified by our pipeline belong to lineages different from the co-cultured symbionts, discarding contamination as a potential source.

To understand the impact of HGT in Asgard evolution, we focused on bacteria-to-Asgard transfers in subsequent analyses. A functional assessment of the transferred genes further supports the lineage-specificity of the transfers, with all enriched GO terms displaying a patchy distribution (**Fig. 3D**). Similar results were obtained when using the high-confidence set, and the same applies for KEGG functional categories (**Fig. 3D**), despite the general KEGG category "Metabolic pathways" being a common enrichment. This is in line with previous observations in prokaryotes as a whole, where ancient transfers have been found to be enriched for genes involved in amino acid, carbohydrate and energy metabolism[20].

ThyX, previously identified as an HGT event in Lokiarchaeia[29], is confirmed in this analysis, as "one carbon pool by folate" is enriched in *Ca.* P. syntrophicum with a 3.25 fold enrichment. This function is also enriched in other close lineages: Helarchaeales and Thorarchaeia (although not significantly after FDR correction), but not significantly enriched in *Ca* L. ossiferum, despite detection of some genes of bacterial origin for the folate pathway in this organism as well.

We next asked whether HGT-derived proteins were integrated in the recipient organism interactome. To do so, we calculated the proteome-wide protein-protein interactions using STRING[30] and assessed the topological properties of the resulting interactome. We found that HGT-derived proteins have a smaller degree, that is, they were less connected than proteins of vertical descent, and were less central in the network (**Figure 3E**). This was true for all lineages.

We found that HGT-derived proteins tend to be more connected to proteins of vertical origin than to other HGT proteins, when normalizing by their respective abundance (**Figure 3E**). This analysis was complemented by manual inspection of the enriched KEGG pathways: in many such pathways, the transferred proteins occupy peripheral steps, interspacing vertically inherited proteins. Therefore, we conclude that HGT does not generally involve the addition of whole modules, but rather a stepwise addition of enzymes to native pathways.

For example, in the aforementioned folate pathway (map00670), despite an overall enrichment in HGT, enzymes encoded by HGT genes and vertically inherited genes seem to be interspaced, sometimes carrying on sequential reactions (**Extended Data Figure 5**). A similar scenario occurs with the carbon metabolism pathway (**Extended Data Figure 6**).

We searched for "HGT islands", that is, segments of the genome in which more than one HGT gene is located contiguous to another HGT gene or interspaced by a single vertically inherited gene. Most of the HGT genes are interspaced with vertically inherited genes, diminishing the likelihood that they are due to errors in the metagenome assembly. However, we found 404 HGT "islands", comprising 901 HGT genes (on average, 33.3% of the HGT genes, with median 31.89%). These islands contained up to 9 genes, with the median being 2. 70 of these islands contained genes belonging to the same COG category, suggesting that they may be operons. We studied these cases in detail. Some of these islands were composed of the same transferred gene duplicated in tandem, such as Coenzyme A transferase or proteins with the B12 binding domain in the Jordarchaeia representative (GCA_019056875.1) (**Extended Data Figure 7**).

An illustrative example of an operon HGT island is that formed by genes encoding the DnaK/DnaJ/grpE chaperone system (HSP70) (**Extended Data Figure 7**) in the Heimdallarchaeial representative (GCA_019057815.1). This system is hypothesized to have first appeared in mesophilic bacteria, and then transferred to archaea by HGT (PMID: 34001607, 23181628, 10048042), an example of HGT extending beyond metabolism to

housekeeping genes.In this Heimdallarchaeial representative the acquired bacterial dnaJ and grpE replaced the original archaeal copies, but not dnaK. The HSP70 system has been linked to adaptation to a mesophilic environment[31] (and Heimdallarchaeia are assumed to be mesophiles[19]), as well as increase in proteome complexity in bacteria[32]. This, however, seems to be a relatively recent adaptation restricted to this specific genome (for grpE) and 9 others within UBA460 (for dnaJ) and may reflect the specific environmental constraints of this species. We observed a total of 10 cases of HGT events of proteins of HSP70 family, both Bacteria-to-Asgard and Asgard-to-Bacteria. Another interesting example is CheA-CheY complex for chemotaxis in Helarchaeales (GCA_011365055.1), CheC-CheD in Hodarchaeales (GCA_016839295.1) (**Extended Data Figure 7**).

Last, we set to assess which HGT proteins may have reached the eukaryotic lineage via the Asgard ancestor of the Last Eukaryotic Common Ancestor (LECA). We assessed all 430 trees flagged as HGT (both Asgard-to-Bacteria and Bacteria-to-Asgard, as we were also interested in assessing whether proteins in eukaryotes may be falsely identified as of bacterial origin) which contained at least one eukaryotic sequence, and 38 contained sequences from at least 3 eukaryotic supergroups suggesting they might have been present in LECA[33]. Upon manual inspection of these candidate phylogenies, we did not find any clear case where a gene of bacterial origin was horizontally transferred to Asgard archaea and passed on through the eukaryotic lineage by vertical inheritance. Interestingly, however, we do find instances where viruses (mainly Nucleocytoviricota, as 25/30 viral sequences found in these trees belong to this clade) seem to act as mediators for the transfer of genes from Asgard archaea to eukaryotes, supporting and expanding previous observations[33].

Discussion

Altogether, our findings depict a highly dynamism in Asgard archaeal genomes and provide the first comprehensive assessment of gene duplication, loss, and HGT within this recently discovered and relevant clade. We reconstructed a large and open Asgard archaeal pangenome, which likely reflects the clade's immense functional and ecological diversity and suggests a vast, untapped genomic diversity. While acknowledging the current dominance of metagenome-assembled genomes (MAGs) and the scarcity of cultured representatives, which may lead to incorrectly inferred absences due to genome incompleteness, we anticipate the pangenome will likely remain open even with a greater diversity of culture-based genomes.

We identified a core set of 53 Asgard signature proteins (ASPs) widespread in the clade and rarely found in prokaryotic out-groups. Significantly, these ASPs are enriched in endocytosis and motor protein functions, two functions that are relevant to eukaryogenesis[34,35]. Notably, several ASPs are shared with eukaryotes and were previously considered Eukaryotic Signature Proteins (ESPs)[2], indicating that Eukaryotes inherited some of the distinctive Asgard archaeal features.

Our comprehensive phylogenomic analysis revealed high gene duplication rates across all lineages—a finding atypical for prokaryotic genomes. Indeed, Asgard archaea exhibited duplication rates intermediate between their closest archaeal relatives (the TACK phylum) and eukaryotic unicellular free-living heterotrophs. This underscores the "stepping-stone" nature of Asgard archaea, even in this genomic trait. Similar to eukaryotes, Asgard archaeal duplications preferentially affected specific gene families, showed lineage-specific patterns, often occurred in tandem, and, alongside gene loss, emerged as a primary driver of genome size. Among the lineages with the highest duplication rates were Heimdallarchaeia (considered the closest sister group to eukaryotes[19]), Lokiarcheia and Sifarchaeia.

We also assessed intra-Asgard transfer, uncovering an intricate network superimposed on a strong vertical inheritance backbone. Intriguingly, genes encoding ESPs were more frequently transferred than other genes, which could explain their previously observed patchy phylogenomic distribution[2]. We estimated that 2.4-14.0% of the genes in Asgard genomes were involved in inter-domain HGT with bacteria. This phenomenon is pervasive across the entire Asgard phylogeny and involves bacterial partners that co-occur with Asgard archaea in metagenomic samples. Inter-domain HGT rates are comparable to those found in TACK (when normalizing for genome size), suggesting HGT does not solely account for increased genome sizes in Asgard archaea. The functions of acquired bacterial genes were highly diverse and lineage-specific, predominantly metabolic, and were found to be peripheral in protein-protein and metabolic networks, interspersed with genes of vertical inheritance This pattern suggests a piecemeal integration into established pathways, perhaps replacing ancestral archaeal proteins, rather than the wholesale addition of new functional modules.

In summary, our work provides a detailed survey of genome content dynamics in Asgard archaea, defines the signature proteins of this important clade, and uncovers a rich dynamism of gene family duplication that is atypical for prokaryotes but reminiscent of processes observed in eukaryotic organisms. These insights offer crucial mechanistic details about the

genomic evolution of this pivotal lineage and its implications for understanding the origins of eukaryotic complexity.

References

1. Koonin, E. V. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol.* **11**, 209 (2010).
2. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
3. Akıl, C. & Robinson, R. C. Genomes of Asgard archaea encode profilins that regulate actin. *Nature* **562**, 439–443 (2018).
4. Rodrigues-Oliveira, T. *et al.* Actin cytoskeleton and complex cell architecture in an Asgard archaeon. *Nature* (2022) doi:10.1038/s41586-022-05550-y.
5. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
6. Mattock, J. & Watson, M. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. *Nat. Methods* **20**, 1170–1173 (2023).
7. Imachi, H. *et al.* Isolation of an archaeon at the prokaryote-eukaryote interface. *Nature* **577**, 519–525 (2020).
8. Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).
9. Inagaki, F. *et al.* Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 2815–2820 (2006).
10. Vetriani, C., Jannasch, H. W., MacGregor, B. J., Stahl, D. A. & Reysenbach, A. L.

Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. *Appl. Environ. Microbiol.* **65**, 4375–4384 (1999).

11. Takai, K. & Horikoshi, K. Genetic diversity of archaea in deep-sea hydrothermal vent environments. *Genetics* **152**, 1285–1297 (1999).

12. Manoharan, L. *et al.* Metagenomes from Coastal Marine Sediments Give Insights into the Ecological Role and Cellular Features of Loki- and Thorarchaeota. *MBio* **10**, (2019).

13. Cai, M. *et al.* Ecological features and global distribution of Asgard archaea. *Sci. Total Environ.* **758**, 143581 (2021).

14. Orsi, W. D. *et al.* Metabolic activity analyses demonstrate that Lokiarchaeon exhibits homoacetogenesis in sulfidic marine sediments. *Nat Microbiol* **5**, 248–255 (2020).

15. Liu, Y. *et al.* Comparative genomic inference suggests mixotrophic lifestyle for Thorarchaeota. *ISME J.* **12**, 1021–1031 (2018).

16. Bulzu, P.-A. *et al.* Casting light on Asgardarchaeota metabolism in a sunlit microoxic niche. *Nat. Microbiol.* **4**, 1129–1137 (2019).

17. Pushkarev, A. *et al.* A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. *Nature* **558**, 595–599 (2018).

18. Wu, F. *et al.* Unique mobile elements and scalable gene flow at the prokaryote-eukaryote boundary revealed by circularized Asgard archaea genomes. *Nat. Microbiol.* **7**, 200–212 (2022).

19. Eme, L. *et al.* Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* **618**, 992–999 (2023).

20. Dmitrijeva, M. *et al.* A global survey of prokaryotic genomes reveals the eco-evolutionary pressures driving horizontal gene transfer. *Nat. Ecol. Evol.* **8**, 986–998 (2024).

21. Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I. & Koonin, E. V. Genomes

in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* **12**, 66 (2014).

22. Treangen, T. J. & Rocha, E. P. C. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet.* **7**, e1001284 (2011).

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

24. Chao, A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43**, 783–791 (1987).

25. Fuentes, D. *et al.* PhylomeDB V5: an expanding repository for genome-wide catalogues of annotated gene phylogenies. *Nucleic Acids Res.* **50**, D1062–D1068 (2022).

26. Morel, B., Williams, T. A., Stamatakis, A. & Szöllősi, G. J. AleRax: a tool for gene and species tree co-estimation and reconciliation under a probabilistic model of gene duplication, transfer, and loss. *Bioinformatics* **40**, (2024).

27. Zhang, J. *et al.* Deep origin of eukaryotes outside Heimdallarchaeia within Asgardarchaeota. *Nature* (2025) doi:10.1038/s41586-025-08955-7.

28. Valentin-Alvarado, L. E. *et al.* Genetic elements and defense systems drive diversification and evolution in Asgard archaea. *bioRxiv* (2024) doi:10.1101/2024.03.22.586370.

29. Filée, J. *et al.* Bacterial origins of thymidylate metabolism in Asgard archaea and Eukarya. *Nat. Commun.* **14**, 838 (2023).

30. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–261 (2003).

31. Rebeaud, M. E. On adaptation to a mesophilic environment and the chaperone network in Archaea. *bioRxiv* (2025) doi:10.1101/2025.06.02.657341.

32. Pan, Z., Zhuo, L., Wan, T.-Y., Chen, R.-Y. & Li, Y.-Z. DnaK duplication and specialization in bacteria correlates with increased proteome complexity. *mSystems* **9**, e0115423 (2024).

33. Bernabeu, M., Manzano-Morales, S., Marcet-Houben, M. & Gabaldón, T. Diverse ancestries reveal complex symbiotic interactions during eukaryogenesis. *bioRxiv* (2024) doi:10.1101/2024.10.14.618062.

34. Yutin, N., Wolf, M. Y., Wolf, Y. I. & Koonin, E. V. The origins of phagocytosis and eukaryogenesis. *Biol. Direct* **4**, 9 (2009).

35. Hoyt, M. A., Hyman, A. A. & Bähler, M. Motor proteins of the eukaryotic cytoskeleton. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 12747–12748 (1997).

Online Methods

Data retrieval and annotation

We compiled a set of 31 genomes, covering the 11 major Asgard lineages (Helarchaeales, Lokiarchaeia, Hermodarchaeia, Thorarchaeia, Hodarchaeales, Heimdallarchaeia, Wukongarchaeia, Sifarchaeia, Odinarchaeia, Baldrarchaeia, and Jordarchaeia) and chosen for their best assembly quality parameters, as well as a random set of equal size and same lineage-based representation for quality control of the pangenomic measures. Taxonomies were assigned *sensu* GTDB, and in cases where one of the lineages was considered a sublineage within another (e.g Hodarchaeales is considered part of Heimdallarchaeia *sensu* GTDB) we ensured that the representative for the higher-level lineage was not part of the low-level lineage (e.g the Heimdallarchaeia representative was chosen from a lineage outside Hodarchaeales).

The genomic sequences and protein predictions for the cultured isolates *Candidatus* Prometheoarchaeum syntrophicum MK-D1[7] (assembly accession GCF_008000775.1) and *Candidatus* Lokiarchaeum ossiferum/Lokiarchaeum sp. B-35[4] (GenBank code CP104013.1) were downloaded from NCBI Assembly and NCBI Nucleotide/Protein, respectively. These genomes served as a baseline, as findings of HGT on Metagenome Assembled Genomes (MAGs) can be contentious due to the possibility of contaminating sequences in the assemblies.

For the MAGs, we selected one best quality (see below) representative for each of the major lineages of Asgard archaea from the Genome Taxonomy Database[5]: Thor-, Heimdall-, Odin, Hel- (part of Lokiarchaeia in GTDB), Hermod-, Sif-, Jord-, Baldr-, and Wukongarchaeia, as well as a member of Hodarchaeales due to its role as sister of eukaryotes[19], which roughly correspond to class-level lineages *sensu* GTDB. As a baseline, we retrieved a selection of one representative per lineage of archaea *sensu* GTDB, attending to the same criteria. These representatives were the genomes with the highest CheckM[23] completeness, and lowest CheckM contamination in case of ties. Additionally, we ran OMArk[36] for additional assessment of contamination. For the metaphylome, we expanded this initial selection to a maximum of three genomes per major lineage when available, following the same selection criteria, except for representatives of Lokiarchaeia, where we prioritized the cultured isolates.

We additionally retrieved representatives of the TACK superphylum ("p__Thermoproteota" *sensu* GTDB) as well as 15 more diverse prokaryotes. For TACK, we selected 2 genomes for each class-level lineage attending to the same criteria. The extra prokaryotic species (5 bacterial and 10 archaeal), were randomly chosen from those available in PhylomeDB[25].We employed the GTDB taxonomy when possible. Since *Candidatus* Lokiarchaeum ossiferum B-35 is not in GTDB as per r207, and therefore does not have a taxonomical classification within the GTDB framework, we classified it using GTDB-Tk[37]. This placed B-35 as d__Archaea;p__Asgardarchaeota;c__Lokiarchaeia;o__CR-4;f__MK-D1;g__Prometheoarchaeum;s__N/A, that is, in the same genus as MK-D1. However, the original publication places them in separate genera, which we reflected here by using an ad-hoc genus name.

When possible, we kept the original genome annotation. Some genomes were not annotated and we annotated them using Prokka v.1.14.6[38] with the flags --kingdom Archaea and --metagenomes. This is the gene annotation program originally used in the two cultured isolates, which we are using as internal control[4,7], as Prokka is a reliable and widely used prokaryotic genome annotator. We adapted the headers of the proteome files, changed the NCBI taxID when the one in the GTDB metadata table was too broad or conflicting, and uploaded them to PhylomeDB[25] via the Phylomizer.

We annotated the Gene Ontology (GO) terms and domains associated with each protein using InterProScan v5.39-77.0[39], and the COG categories and KEGG KOs with eggNOG-mapper v2[40]. To perform functional assessments across the Asgard pangenome, we also annotated the orthogroups by compiling the annotations of their member proteins following the pipeline in[33].

Last, we taxonomically downsampled the list of FLUH proteomes from Bernabeu et al.[33], randomly removing genomes while maintaining the relative proportions of the different supergroups, keeping a total of 30.

The list of genomes, along with associated metadata, can be found in **Extended Data Table 1**.

Alternative species tree reconstruction

We inferred branch lengths of the Asgard species tree by retrieving the markers for all archaeal species from GTDB, aligning them with MAFFT v7.407[41] (–auto), trimming with trimAl v1.4.rev15[42], and reconstructing the species tree with IQ-TREE v1.6.9[43] (--models

DCmut,JTTDCMut,LG,WAG,VT), both constraining the topology to the current consensus[19] or leaving it unconstrained, resulting in an alternative topology.

Asgard archaeal pangenome assessment

We performed an orthogroup clustering with OrthoFinder v.2.5.4[44] using BLAST[45] as an aligner. To assess pangenome metrics, such as Chao's lower bound, Heaps estimate, and the binomial mixture models we employed the R package micropan[46].

We defined Asgard signature proteins (ASPs) as those OG common (present in ≥ 70% of the proteomes) in Asgard archaea (present in ≥ 70% of the proteomes) but rare (present in ≤15% of the proteomes) in other archaea. We mapped ASPs onto the Protein-Protein Interaction Network (PPI), obtained with STRING[30], of the seed genome with the highest number of ASPs (the representative of Hermodarchaeia, with 50/53).

To perform functional enrichments, we collapsed the annotations at orthogroup level following the pipeline in Bernabeu et al[33].

Genome dynamics across Asgard archaea

We constructed a metaphylome consisting of a collection of phylomes employing one seed per major lineage. We selected the highest quality proteome (by CheckM completeness/contamination) per each major lineage, except in the Loki lineage where we employed the two isolates as seeds. The phylome reconstruction pipeline was applied as defined elsewhere[25]. In brief, for each of the genes contained within each of the seed genomes, we performed a protein BLAST[45] search against all proteomic data to retrieve proteins with significant similarity (defined as e-value < 1e-05, continuous overlap over 50% of the query sequence). We then took the top 150 hits per protein and generated an MSA with three different methods (MUSCLE V3.8.1551[47], MAFFT v7.407[48] and KALIGN v2.04[49]), in forward and reverse, and combined the resulting six alignments into one consensus alignment with M-Coffee v12.0[50]. This alignment was then trimmed with trimAl v1.4.rev15[42], with a consistency cut-off of 0.1667 and gap score cut-off of 0.1. These alignments were then used to reconstruct the phylogenetic tree of each protein family with IQ-Tree v1.6.9[43], for which we selected the final Maximum Likelihood (ML) tree using the best model selected based on the Bayesian Information Criterion (BIC) and support was calculated with ultrafast bootstrap under 1000 repetitions.

We then ran the species overlap algorithm following the PhylomeDB pipeline[25]. We employed both topologies for the prokaryotic dataset and the species tree as inferred by Duptree[51] for the eukaryotic set.

For the reconciliation algorithm, which we ran for the prokaryotic set as we infer HGT to play a significant role in its genome dynamics, we selected one sequence per OG present in the Asgard clade (the longest sequence per orthogroup) and ran AleRax[26] v1.1.0 using as input the distribution of bootstrap trees (as supplied by IQ-TREE) and the generated species tree as backbone, estimating rates per family. We ran AleRax with both models (DL and DTL) and performed a log-likelihood ratio test (p.value = 1), therefore we used the DTL model for the analysis.

Inter-domain HGT assessment

We performed a similarity search with BLAST 2.11.0[45] of the proteomes against a custom-made database (BROADDB_V2), described in[33], which comprises all species representatives from GTDB r207[5], viral sequences from RVDB 25.0[52], and eukaryotes from UniRef v2023_01[53] and EukProtV1[54]. We chose not to employ NCBI NR as it does not provide a sufficient representation of the Asgard genomes. We employed GTDB taxonomy for prokaryotes as it allows a more comprehensive, robust and fine-grained taxonomic analysis as compared to NCBI taxonomy.

We analyzed the BLAST results with HGTector v2.0b3[55], which systematically looks for hit distribution patterns incongruent with vertical evolution, given a series of hierarchically defined evolutionary categories. We ran the algorithm setting the “self” group as Asgard archaea and the “close” as Archaea, therefore allowing us to focus on detection of inter-domain HGT events. The output statistics can be found in **Extended Data Table 3**.

For candidate transfers selected by HGTector, we retrieved the best 300 BLAST hits with an e-value cut-off of 0.001 and used them to reconstruct a gene tree following the PhylomeDB V5 pipeline described before.

When the selected set of homologous sequences was saturated with eukaryotic sequences (that is, if they comprised >80% of the non-Asgard sequences) or Asgard sequences, we performed a subsampling of these clades by selecting the first hit per major eukaryotic group as defined in EukProt taxonomical framework[54], then retrieved all hits from those organisms that had a lower e-value than the lowest e-value for a non-Asgard prokaryotic sequence for

that query. This allows us to have taxonomically-informative insight into the distribution of the sequences across eukaryotes whilst allowing us to analyze the distribution of more distant prokaryotic homologues.

We further analyzed the resulting gene trees with Abaccus v1.2[56], which identifies taxonomical "jumps" within gene trees that do not follow the species tree and therefore further helps discern putative HGT events. For this, we used the taxonomy *sensu* GTDB for prokaryotes.

Additionally, we retrieved the support of the branch of the transfer (e.g. the branch subtending the acceptor Asgard clade and the putative non-archaeal donor clade). We only kept cases when this support was ≥50%, or, if the node was poorly supported, cases where the sister of the transfer branch contained ≤ 10% of archaeal sequences were also considered, even if the specific donor clade could not be ascertained. To ensure that branch support was not affecting the conclusions of the analysis, we built a subset of well-supported (≥ 70% support) HGT cases in which all analyses were also run.

Last, we manually curated the HGT results by filtering out incongruencies and false positives resulting from trees saturated with eukaryotic sequences. We performed this step for all proteins flagged by HGTector, whether they were picked up by Abaccus or not.

To do so, we first numbered the internal nodes in the Asgard tree . Then, we parsed the gene trees with an in-house script (HGT_detection.py) based on ete3[57], using the following criteria:

1. Trees were rooted at midpoint.

2. The start node was established as the one subtending the largest Asgard monophyletic clade containing the seed protein. In assessing monophyly, we considered eukaryotes as a *de facto* Asgard lineage stemming from Heimdallarchaea. The most recent common ancestor (MRCA) of the Asgard lineages present in the clade is assumed as the acceptor Asgard lineage.

3. Then, the sister clade of the start node was analyzed:

    a. If the fraction of non-Asgard Archaeal sequences in the sister clade was ≥ 10%, the tree was labeled as a False Positive. This means that either vertical inheritance cannot be ruled out, or that an Archaeal-to-Asgard HGT cannot be

ruled out, which is also beyond the scope of this analysis since we are focusing on inter-domain HGT.

b. The donor was set as the MRCA of at least 50% of the non-Asgard sequences of the sister clade. In cases where the MRCA is a broad category (e.g "Bacteria"), we established the predominant phyla as the putative donor lineage.

4. Additionally, taxonomy information of the sister clade to the (start+sister) branch (sister2) was retrieved.

5. We assessed the percentage of non-Asgard sequences in the tree that were of eukaryotic origin. In cases where this percentage ≥ 80% were labeled Eukaryotic SIgnature Proteins (ESPs) and submitted to the subsampling mentioned above and the subsampled gene tree was re-analyzed.

After this analysis, we additionally flagged as FP events for which support was low (≤ 50) and sister2 contained ≥ 10% non-Asgard archaeal sequences. These criteria are very restrictive and likely result in an underestimation of the true number of HGT events: namely, we will miss inter-domain HGT and the 10% of Archaeal sequences in the sister needed to label a tree as a false positive will lead to true HGTs being missed due to secondary transfers of the gene family from bacteria to other archaea. However, this ensures that the trees that are picked up as HGT are true HGT events.

To assess the directionality of the transfer (Asgard-to-bacteria or bacteria-to-Asgard, we assessed whether there were Asgard sequences in sister2, if there were, Asgard-to-bacteria was assumed.

Tandem duplications and HGT islands

To infer tandem duplications we used a window-based approach to query the GFF files where we looked for 2 or more genes belonging to the same OG being either directly adjacent or interspaced by another gene, which we applied with a custom R script (05_Tandem_Duplications.R). For HGT islands, we also queried the GFFs to find HGT genes gene immediately adjacent or interspaced by a vertically inherited gene to reconstruct "islands". We visualized it with the R package gggenes 0.5.1 (https://github.com/wilkox/gggenes/tree/master).

Functional enrichment analysis

We performed GO term enrichment analyses using the Bioconductor package topGO (https://git.bioconductor.org/packages/topGO), with the method "weight01" and using the Fisher's exact test, correcting for multiple testing and keeping results with a False Discovery Rate (FDR ≤ 0.1). For KEGG pathway enrichment, we employed the package ClusterProfiler[58] with the same statistical criteria. We visualized the results with the R package ggpubr (https://github.com/kassambara/ggpubr/) and ggtree/ggtreeExtra[59,60].

Protein-protein interaction networks

We further analyzed the functional relationship between transferred genes by assessing the network of protein-protein interactions (PPI). Since none of the analyzed Asgard genomes are available in the database, we first submitted them to be added to STRING[30] by making use of the tool developed in the latest version[61]. We then uploaded the sequences of the transferred genes to the web server and ran the analysis using the newly uploaded proteome as target organism. As Asgard is not a selectable template, we employed Thermoprotei, as it is the closest group that was available for use as template in STRING. We downloaded the results, performed analysis with the R implementation of iGraph (https://github.com/igraph/rigraph) and visualized the networks with the R package ggnetwork (https://journal.r-project.org/archive/2017/RJ-2017-023/index.html).

Code availability

All described scripts are available on Github (https://github.com/Gabaldonlab/Asgard).

Data availability

input proteomes and output files for all steps can be found in Zenodo (https://zenodo.org/records/15789115?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6Ijc5OWY3YjZkLWJlNWYtNGY0My1hYmViLTQyM2VlNmExYzRjMCIsImRhdGEiOnt9LCJyYW5kb20iOiI5YjQzYjUyMzU5NTI3NWFjNWZiZGEyYzU0OGQ1ZTFjYyJ9.MVQ7lZimHu4O9lK2xWsxnzDqwd_SbiAzPKp_WKO2E_GGkAFFtlKEdk_M72p_5leISqCqfN6ZuH-TMWs8aOnb2Q)

References

36. Altenhoff, A. M. *et al.* OMA orthology in 2024: improved prokaryote coverage,

ancestral and extant GO enrichment, a revamped synteny viewer and more in the OMA Ecosystem. *Nucleic Acids Res.* **52**, D513–D521 (2024).

37. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* **38**, 5315–5316 (2022).

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

39. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).

40. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).

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

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

43. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

44. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

45. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

46. Snipen, L. & Liland, K. H. micropan: an R-package for microbial pan-genomics. *BMC*

Bioinformatics **16**, 79 (2015).

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

48. Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

49. Lassmann, T. & Sonnhammer, E. L. L. Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **6**, 298 (2005).

50. Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699 (2006).

51. Wehe, A., Bansal, M. S., Burleigh, J. G. & Eulenstein, O. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* **24**, 1540–1541 (2008).

52. Goodacre, N., Aljanahi, A., Nandakumar, S., Mikailov, M. & Khan, A. S. A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere* **3**, (2018).

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

54. Richter, D. J. *et al.* EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community J.* **2**, (2022).

55. Zhu, Q., Kosoy, M. & Dittmar, K. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics* **15**, 717 (2014).

56. Naranjo-Ortíz, M. A. *et al.* Widespread Inter- and Intra-Domain Horizontal Gene

Transfer of d-Amino Acid Metabolism Enzymes in Eukaryotes. *Front. Microbiol.* **7**, 2001 (2016).

57. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

58. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).

59. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinformatics* **69**, e96 (2020).

60. Xu, S. *et al.* GgtreeExtra: Compact visualization of richly annotated phylogenetic data. *Mol. Biol. Evol.* **38**, 4039–4042 (2021).

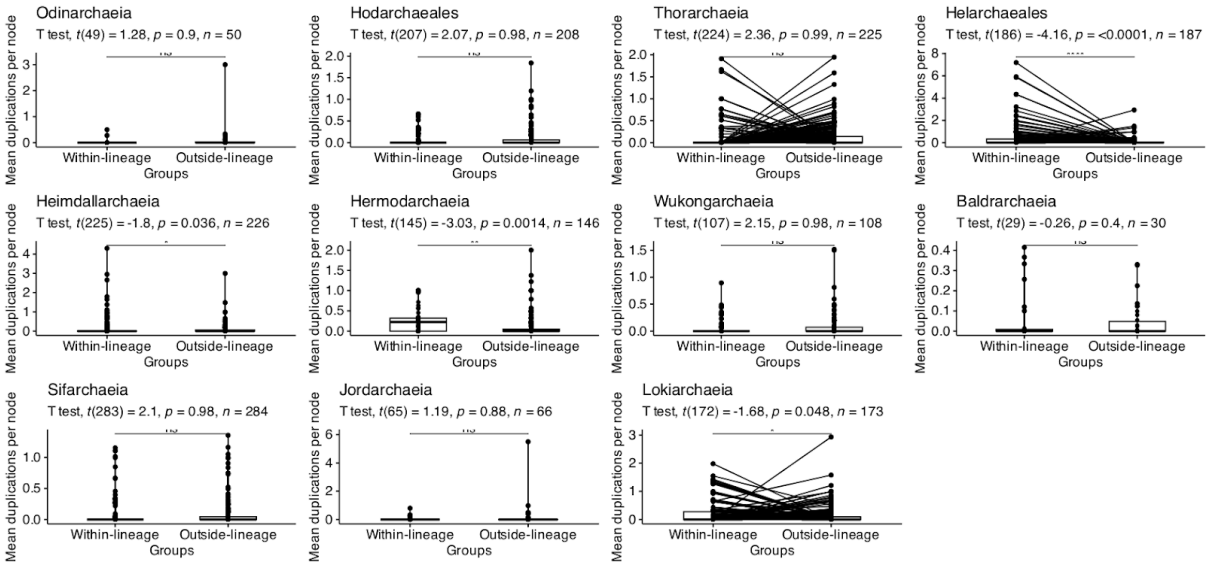
61. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

Extended Data

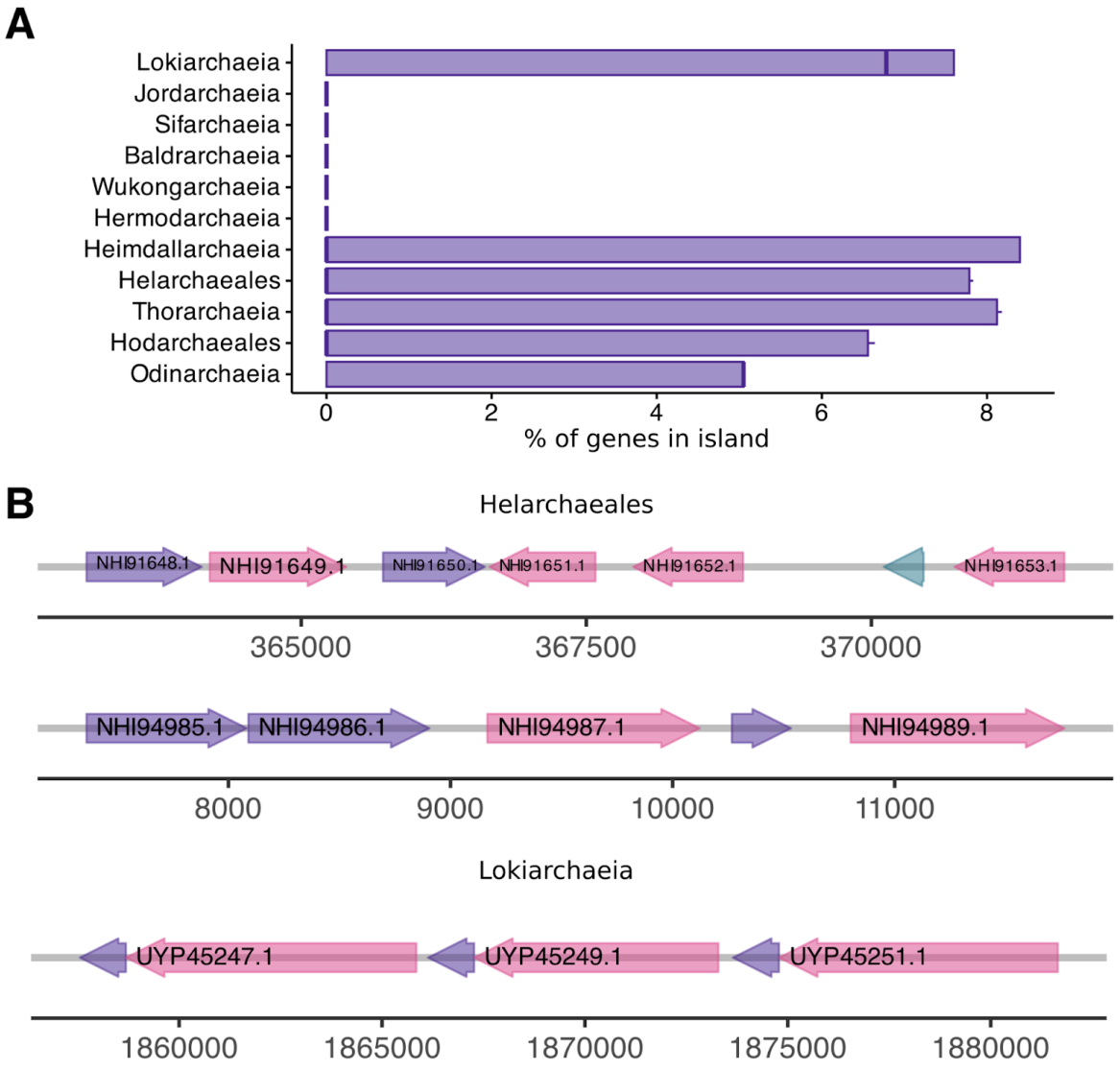
Extended Data Table 1: List of genomes considered for the analyses (prokaryotes and eukaryotes) and associated metadata obtained from GTDB (for prokaryotes).

Extended Data Table 2: KEGG patways enriched per subcategory

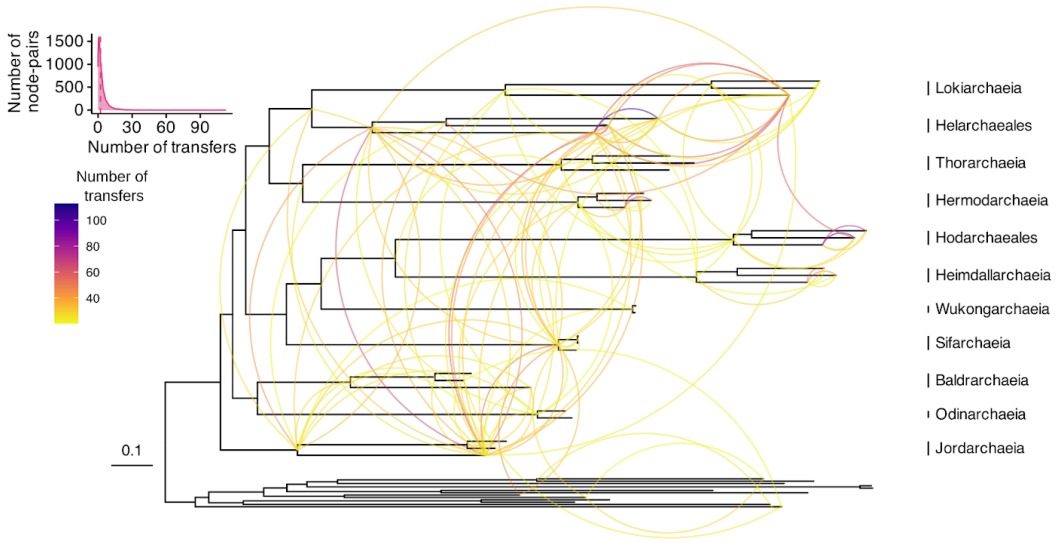
Extended Data Table 3: Main output statistics for the DTL and HGT pipeline for the 12 seed genomes



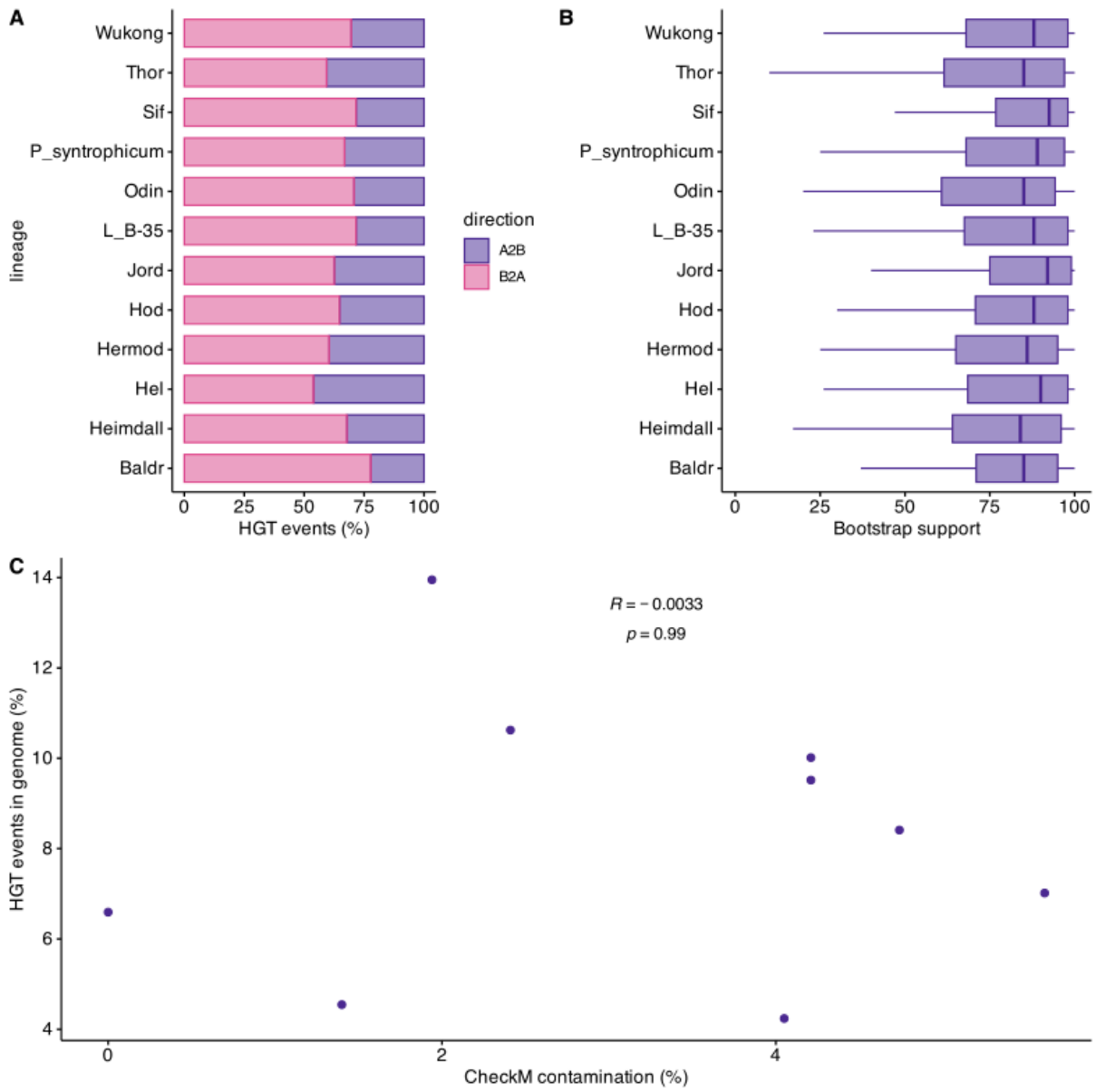
Extended Data Figure 1: Comparison of duplication rates within- and outside-lineage for signature gene families. Signature gene families are defined as those common within the lineage (≥70%) but rare (≤ 15%) outside it. Statistical significance as defined by pairwise t-test.



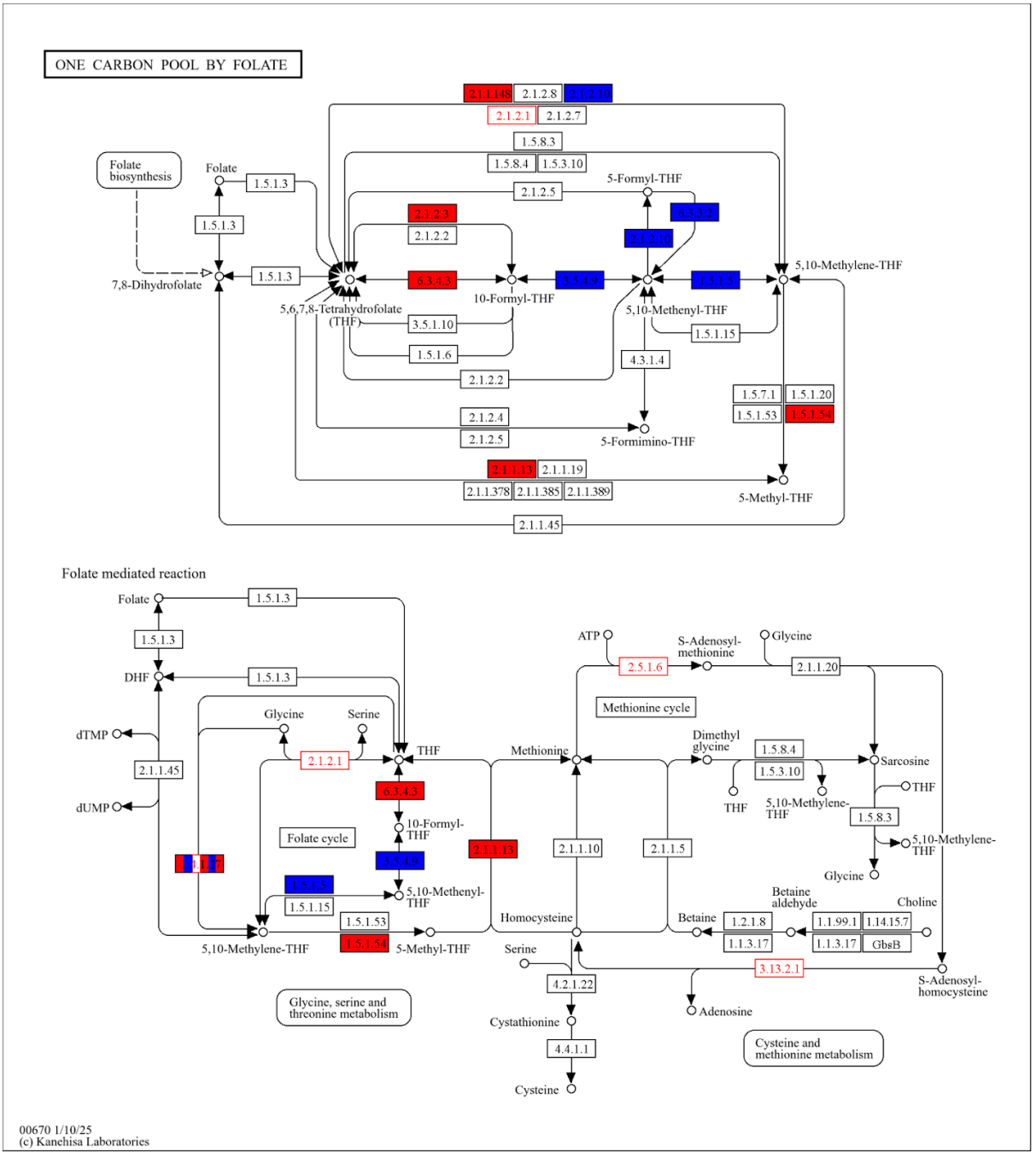
Extended Data Figure 2: Tandem duplications in Asgard archaea (A) Percentage of genes within tandem duplications with respect to the total number of duplicated genes per lineage. (B) Schematic representation of selected cases of tandem duplications. The lineage is depicted to the right of tandem duplication. Within each genomic region, the X axis represents the genomic coordinates. Arrows depict genes and their coding sense.. Within each island, genes with the same color belong to the same orthogroup.



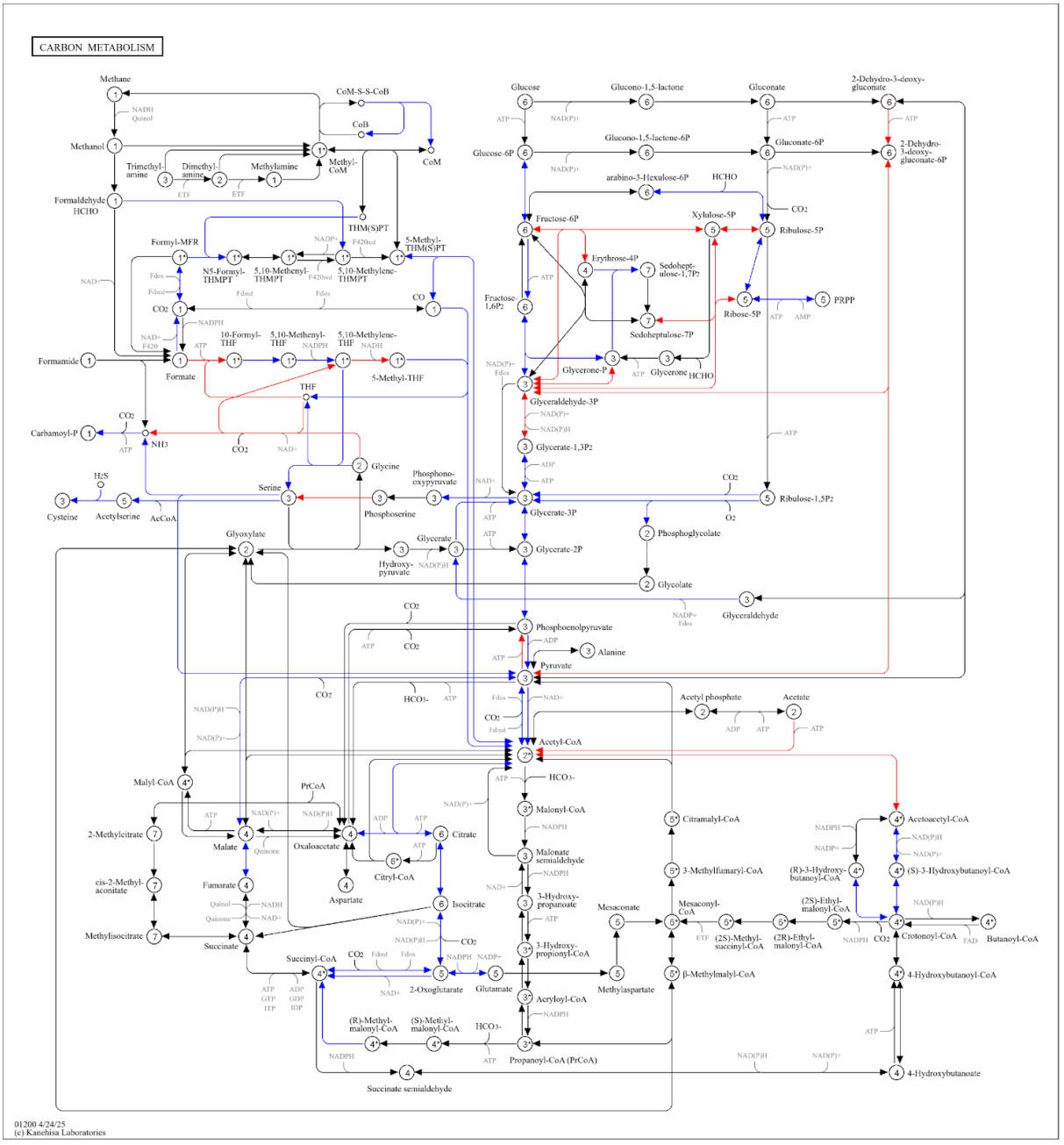
Extended Data Figure 3: Network of transfer pairs as inferred by ALERax against the backbone of the archaeal tree of life. Scale depicts number of substitutions per site. Only node-pairs with a number of transfer events higher than percentile 95 (20) are displayed, distribution of the number of transfers per transfer-pairs depicted in the inset plot on the top right.



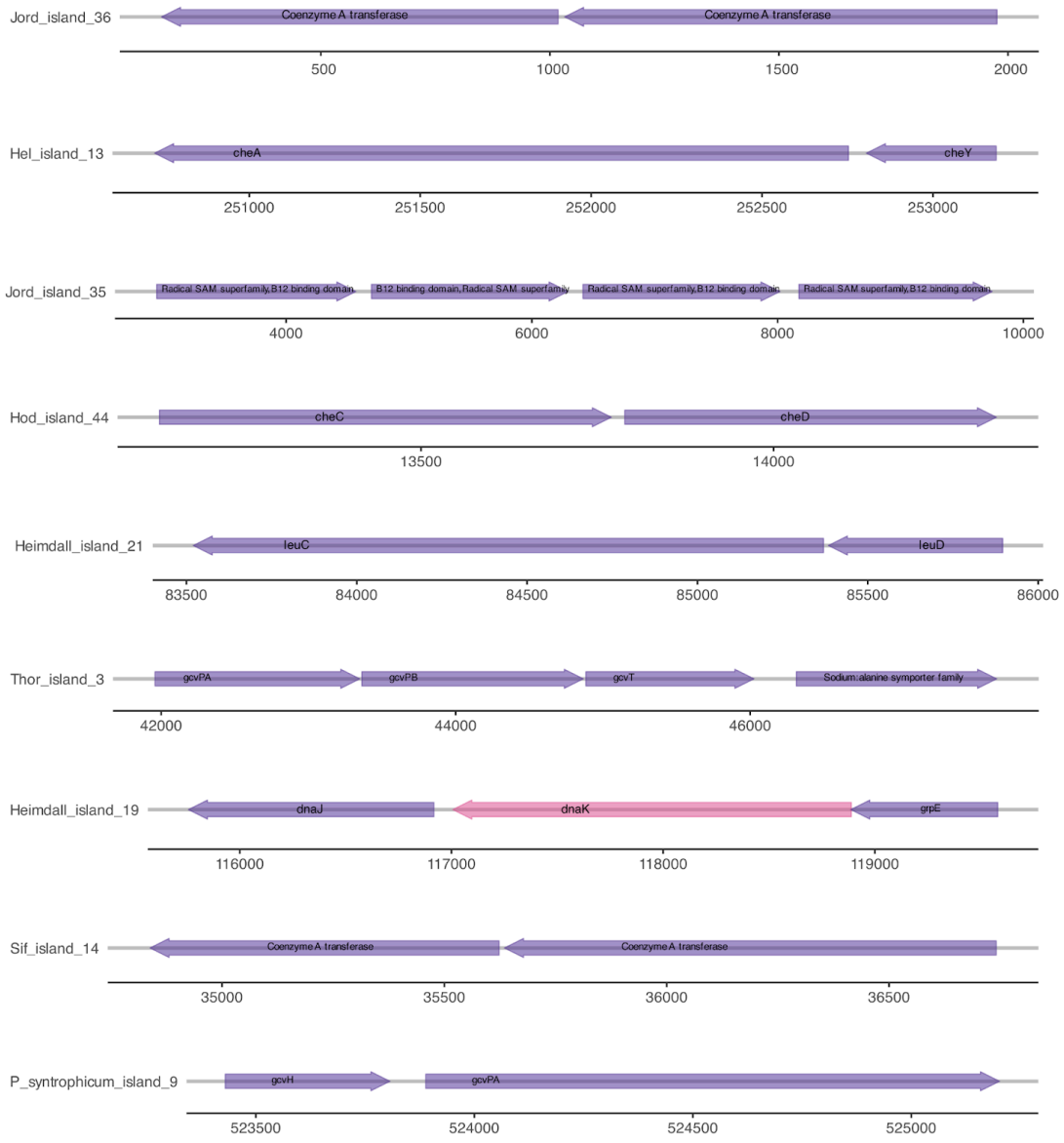
Extended Data Figure 4: (A) Directionality of the transfers. Transfers are considered "Asgard to Bacteria" or A2B if the second sister is majoritary of Asgardarchaeotal taxonomic affiliation, and "Bacteria to Asgard" or B2A if the second sister is majoritary bacterial. (B) Distribution of bootstrap support for the node that joins the Asgard acceptor clade with its inferred sister. (C) Correlation between estimated percentage of HGT genes and estimated CheckM contamination per proteome.



Extended Data Figure 5: Genes associated to the folate pathway in *P. syntrophicum*. One carbon pool by folate (map 00670) depicted. Genes inferred to be HGT shown in red, genes inferred to have been inferred vertically shown in blue.



Extended Data Figure 6: Genes associated to the carbon metabolism pathway in *P. syntrophicum*. One carbon pool by folate (map 00670) depicted. Genes inferred to be HGT shown in red, genes inferred to have been inferred vertically shown in blue.



Extended Data Figure 7: Representative examples of HGT islands. X axis displays genomic coordinates, HGT genes are marked in purple. The vertically inherited gene is indicated in pink.