Viruses associated with *Nitrososphaerota* archaea: diversity, evolution and horizontal gene transfers

Kaiyang Zheng[1], Yantao Liang[1,2,3,4,5,6,7*], David Paez-Espino[8], Andrew McMinn[1,4,7,9*], Min Wang[1,2,3,4,5,6,7,10*]

[1] College of Marine Life Sciences, Ocean University of China, Qingdao, China.

[2]Haide College, Ocean University of China, Qingdao, China.

[3]MoE Key Laboratory of Evolution & Marine Biodiversity, Ocean University of China, Qingdao, China.

[4] Institute of Evolution and Marine Biodiversity, Ocean University of China, Qingdao, China.

[5] Frontiers Science Center for Deep Ocean Multispheres and Earth System, Ocean University of China, Qingdao, China.

[6] Center for Ocean Carbon Neutrality, Ocean University of China, Qingdao, China.

[7] UMT-OUC Joint Centre for Marine Studies, Qingdao, China.

[8]Ancilia Biosciences, Co. Ltd., Alexandria Launchlabs, New York, USA.

[9] Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, TAS, Australia.

[10] The Affiliated Hospital of Qingdao University, Qingdao, China.

* Correspondence:

Yantao Liang, Email: liangyantao@ouc.edu.cn

Andrew McMinn, Email: andrew.mcminn@utas.edu.au

Min Wang, Email: mingwang@ouc.edu.cn

Abstract

Nitrososphaerota, an archaeal phylum mainly containing *Thaumarchaea* and *Aigarchaea*, is an ecological widespread archaeal lineage belonging to TACK super-phylum. However, viral diversity and evolution associated with this archaeal lineage, especially for extremophilic *Thaumarchaea* and *Aigarchaea*, remain uncharacterized. Here, we reveal the neglected diversity of the *Thaumarchaea* virome by identifying 152 species-level viral and mobile genetic element (MGE) sequences associated with extremophilic *Thaumarchaea*, which mainly colonize in hyperthermophilic environments, deep sea, and acidic soils. Importantly, we identified 263 previously unrecognized species-level viral/MGE sequences associated with *Aigarchaea*, including a novel rod-shaped viral family uniquely characterized by capsid heterodimers, thereby filling a long-standing gap in the viral diversity of the TACK super-phylum. Phylogenetic analysis revealed close evolutionary relationships between *Aigarchaea*-associated rod-shaped viruses and *Methanophagales* (ANME-1) rod-shaped viruses, suggesting that they may have evolved from a common rod-shaped viral ancestor. In contrast, *Aigarchaea*-associated head-tailed viruses formed three distinct monophyletic clades compared to other viruses, representing previously undiscovered evolutionary lineages within virosphere. Gene functional analysis identified 32 viral-encoded genes potentially involved in supplementing host metabolism, including, for the first time, genes that confer selective advantages to their extremophilic archaeal hosts, such as those involved in oxidative stress adaptation and antibiotic synthesis. Approximately 54% of 6,505 horizontal gene transfers (HGTs) occur bidirectionally between *Nitrososphaerota* and viruses/MGEs, highlight the significant role of viruses/MGEs in facilitating gene exchange within the archaeal communities. Collectively, this work sheds light on the previously unknown *Aigarchaea*-associated viromes, and unveils the underexplored diversity, evolution, functional repertoires, and HGTs of *Nitrososphearota*-associated viruses.

Introduction

The archaea phylum *Nitrososphaerota* within TACK super-phylum mainly contains *Thaumarchaea* and *Aigarchaea* [1,2]. In aquatic ecosystems, *Thaumarchaea* constitute up to 40% of microorganisms, contributing significantly to ammonia oxidation, thus heavily influencing nitrogen cycling [3-5]. *Thaumarchaea* including all ammonia-oxidizing archaea (AOA) [6], utilize ammonia oxidation pathways as their primary energy source [7,8]. Compared to ammonia-oxidizing bacteria, AOA thrive in low-ammonia conditions due to higher ammonium affinity, dominating nitrogen-limited environments as efficient "nitrogen thieves" [9-14]. Originating from terrestrial, non-ammonia-oxidizing ancestors, *Thaumarchaea* is thought to have adapted aerobically post-Great Oxygenation Event, becoming abundant in diverse environments and contributing substantially to carbon fixation and ammonia oxidation, even in extreme habitats [8,15-20]. As the sister clade of *Thaumarchaea*, *Aigarchaea* shares substantial genetic content involved in sulphur oxidation and amino oxidation, suggesting an ancestral ecological transition from thermal to moderate environments [21]. Primarily inhabiting hydrothermal vents and hot springs, *Aigarchaea* influence carbon and sulphur cycling, nutrient transformation, and microbial community structures, playing critical ecological roles in geothermal systems [22,23].

Viruses and mobile genetic elements (MGEs) significantly shape archaeal genomic and ecological landscapes, exhibiting high-level of morphological and genomic diversity relative to viruses from other life domains [24]. Due to cultivation challenges, environmental metagenomics has become the primary method for exploring archaeal viral diversity. Recent studies have greatly expanded viral diversity associated with the Asgard super-phylum, ANME-1, and DPANN super-phylum, addressing gaps in viral evolution and host-virus interactions [25-29]. Over 50 *Thaumarchaea*-associated viral/MGE sequences from marine samples have so far been identified, revealing diverse viral elements that influence ammonia oxidation

and host genomic integrity [30-36]. However, the limited number of case studies prevents a comprehensive profiling of the genomic and functional diversity of these archaeal viruses. In particular, the knowledge gap of viruses associated with extremophilic *Thaumarchaea* lineages (e.g., *Nitrosocaldales* and *Nitrosotalea*) poses a significant challenge for systematically resolving their phylogenetic placements within the virosphere. Furthermore, as a phylogenetic sister clade closely related to *Thaumarchaea* and one of the main lineages constituting the TACK super-phylum, *Aigarchaea*-associated viruses/MEGs remain entirely underexplored to date, representing a major enigma within virology.

To address this knowledge gap, we systematically explored *Thaumarchaea* and *Aigarchaea* (collectively referred to as *Nitrososphaerota* in this study) associated viral/MGE diversity using extensive global metagenomic datasets. Unlike previous studies [30-34], our analysis of extensive global datasets enabled detailed taxonomy, phylogeny, metabolic potential analyses, and host-virus interactions. Our research established a genomic catalogue comprising 746 viral/MGE sequences (624 species-level sequences) associated with *Nitrososphaerota*, spanning diverse extreme and moderate ecosystems globally, and shedding light on the previously unknown *Aigarchaea*-associated viromes.

Results

Overview of the genomic catalogue of *Nitrososphaerota*-associated viruses/MGEs

The mining of *Nitrososphaerota*-associated viruses/MGEs primarily utilized public viral databases, *Nitrososphaerota* genomes, and metagenomic samples from hydrothermal and deep-sea environments (Fig. 1a and Supplementary Note 1). A comprehensive *Nitrososphaerota* genome database containing 2,137 genomes and metagenome-assembled genomes (MAGs) was constructed from NCBI GenBank and IMG/M databases [37,38]. Metagenomic datasets from extreme habitats, including hyperthermal, abyssal/hadal ocean, hypersaline, frozen, and acidic environments, were analyzed (Supplementary Dataset 1) [37,39-41]. Employing a hybrid identification strategy integrating CRISPR spacer matches and protein-sharing analysis, a total of 746 *Nitrososphaerota*-associated viruses/MGEs were identified, including 712 extrachromosomal viruses/MGEs, and 34 integrated viruses/MGEs. These were grouped into 624 species-level viral operative taxonomic units (vOTUs), comprising 153 vOTUs with known viral capsid proteins and 471 vOTUs lacking annotated capsids (Fig. 1a, Supplementary Dataset 2). The pipeline prioritized specificity, verifying 14 previously reported viruses/MGEs [31,32] (Fig. 1b), and newly identified 350 viral/MGE contigs (336 vOTUs) that were associated with *Nitrososphaerota*. Most viruses/MGEs originated from geothermal, marine, soil, and hydrothermal environments, aligning with *Thaumarchaea* and *Aigarchaea* habitats [7,19,22].

Identifying the host and habitat diversity of *Nitrososphaerota*-associated viruses/MGEs

The 1,192 CRISPR-Cas cassettes were identified across 636 *Nitrososphaerota* genomes (Supplementary Dataset 3) and these were used to describe associations between *Nitrososphaerota* and viruses/MGEs. These cassettes were classified into five distinct CRISPR types, whose spacer diversity reflected more extensive diversity

of *Nitrososphaerota* viromes. The spacer-sharing network has low viral coverage (only 336 viral/MGE sequences) (Supplementary Dataset 4), suggesting broader undiscovered viral diversity. Host-virus specificity varied among *Nitrososphaerota* genera: for example, spacers from *Caldarchaeum* demonstrated genus-level specificity, whereas other *Thaumarchaea* genera indicated more strict viral specificity (Fig. 2a). In contrast, spacers from *Aigarchaea* lineage, such as *Terraquivivens* and *Wolframiiraptoraceae*, showed broader sharing patterns, suggesting weaker genus-level specificity.

Nitrososphaerota-associated viruses/MGEs were retrieved from diverse ecosystems and linked to multiple archaeal lineages. They are widely distributed from tropical to polar regions and across terrestrial and marine environments (Fig. 2b). With regard to habitat, Group 1.1a and MT/GW (marine trench and ground water) -associated viruses/MGEs were mainly found in aquatic environments, while Group 1.1b-associated viruses/MGEs are predominantly from soil (Fig. 2c). Viruses/MGEs from geothermal and hydrothermal environments were primarily linked to *Caldarchaeales* and *Wolframiiraptoraceae*, aligning with their archaeal hosts' habitat preferences [21,42-45]. Of the host lineages, 746 of the identified viruses/MGEs spanned 11 *Nitrososphaerota* lineages (Fig. 2d and Supplementary Dataset 2) [2,20,21,23,46]. For the 352 *Thaumarchaea*-associated vOTUs, 287 linked to AOA, while 65 were associated with Non-AOA. These viruses/MGEs included 152 previously undescribed vOTUs associated with extremophilic *Thaumarchaea* lineages, comprising 63, 51, 15, 12, and 12 vOTUs related to MT/GW, Deeply-branching *Thaumarchaea* (D-B Thaum), Marine Beneath Group A (MBG-A), *Nitrosotalea*, and *Nitrosocaldales*, respectively. Of the 263 newly identified *Aigarchaea*-associated vOTUs, 210 linked to *Wolframiiraptoraceae*, 49 were linked to *Caldarchaeales*, and four were linked to pSL4-affiliation. With regard to morphology, head-tailed viruses, known as cosmopolitan prokaryotic viruses, are widespread across 11 *Nitrososphaerota* lineages (n = 447 vOTUs) (Fig. 2d and Supplementary Dataset 2), aligning with prior studies [25-27,30,32-34,47]. Additionally, 44 previously undescribed *Aigarchaea*-associated rod-shaped vOTUs were identified. These findings show that viruses/MGEs linked to nine

additional archaeal lineages in contrast to previous studies [30-36,48].

Genomic and phylogenetic diversity of viruses/MGEs associated with *Nitrososphaerota*

Genome-content-based analysis was used to characterize relationships and diversity among *Nitrososphaerota*-associated viruses/MGEs. Among these, 16 viral/MGE genomes (six viral clusters) showed genome-wide linkages to other bacterial and archaeal viruses (Supplementary Fig. 2 and Supplementary Dataset 5); ten genomes shared significant genomic similarity (>20% shared protein clusters) with established archaeal viral families (*Chaacviridae*, *Ekchuahviridae*, and *Itzamnaviridae*), which typically infect methanogenic archaea with diverse morphologies [27,28]. However, most *Nitrososphaerota*-associated viruses/MGEs (95.2%) exhibit unique genome content, forming 119 independent viral clusters (Supplementary Fig. 2). These clusters belong to 58 groups based on ≥20% protein-cluster sharing, including 31 *Thaumarchaea*-specific and 25 *Aigarchaea*-specific groups (Fig. 3a). These reflect host lineage and morphological specificity consistent with previous studies [27,47].

Twelve families encoding known capsid proteins were found to be associated with *Nitrososphaerota*, including eight previously undescribed families. Of the established viral families, three families (*Ekchuahviridae*, *Itzamnaviridae*, *Chaacviridae*), initially identified from deep-sea hydrothermal vents linked to ANME-1 archaea [28], were found to be also associated with *Aigarchaea* (*Wolframiiraptoraceae*) (Supplementary Dataset 2 and Fig. 3b). Of the possible new viral families, one rod-shaped family (*Byatisviridae*) and four head-tailed families (*Fthagguaviridae*, *Sthaneeviridae*, *Ythogthaviridae*, and *Cthughaviridae*) were established, which were associated with *Aigarchaea* (*Wolframiiraptoraceae* and *Caldarchaeales*); two head-tailed families (*Ommogviridae* and *Yebviridae*) were established and these were associated with *Thaumarchaea*. Notably, *Ommogviridae* was initially identified in a previous metagenomic study, characterized by their nitrification genes (ammonia monooxygenase, *amoC*) [33]. Moreover, 18 MGE groups

which lacked capsid proteins were detected, forming independent groups in the genome-content-based network. These suggest that they have unique genomic features compared with other reference viral genomes (Supplementary Dataset 2, Supplementary Figs. 2 and 3).

This comprehensive genomic catalogue of *Nitrososphaerota*-associated viruses/MGEs enables systematic assessment of their phylogenetic placements using Bayesian (BI) and Maximum-Likelihood (ML) joint phylogenetic inference (Extended Data 3 and 4). For *Nitrososphaerota*-associated rod-shaped viruses (*Byatisviridae*), two genomes showed close relationships with ANME-1-associated *Ahmunviridae* viruses [28]. *Byatisviridae* members formed a distinct monophyletic sister clade to *Ahmunviridae*, suggesting divergence from a common ancestor (Fig. 4a). This is supported by the unique structural characteristics of their SIRV2-like fold capsid proteins (Supplementary Fig. 5a). Unlike *Rudiviridae* and *Ahmunviridae*, which share a common evolutionary origin, *Byatisviridae* appears to possess a capsid composed of heterodimers, implying that those heterodimeric capsids of *Aigarchaea*-associated rod-shaped viruses may have evolved from an ancestral homodimeric architecture (Fig. 4a, Supplementary Figs. 5b and 6). For head-tailed viruses, most *Aigarchaea*-associated viruses formed monophyletic clades in phylogenetic trees. Some *Thaumarchaea*-associated viruses are phylogenetically intertwined with other archaeal and bacterial viruses. They exhibit typical polyphyletic characteristics. *Fthagguaviridae* and *Cthughaviridae* are two *Aigarchaea*-associated viral families, forming two distinct monophyletic groups in the phylogenetic trees of capsid and terminase, respectively, which is further supported by capsid structural similarity networks (Figs. 4b, 4e, and Supplementary Fig. 7). This indicates that these viruses occupy entirely distinct evolutionary placements from other archaeal viruses and bacteriophages. *Ekchuahviridae* clustered within archaea-associated viruses (*Ahpuchviridae*, *Halomagnusviridae*, and *Haloferuviridae*), and is closely related to bacteriophages infecting *Actinomycetes* (Fig. 4c). *Yebviridae* likely shares a common ancestor with T4-like cyanophages (*Kyanoviridae*); and this is supported by capsid structural similarity networks (Figs. 4e, 4f, and Supplementary Fig. 7). Collectively,

an additional seven *Nitrososphaerota*-associated viral families are proposed and the evolutionary placement of these archaeal viruses is elucidated.

Metabolic potential of *Nitrososphaerota*-associated viruses/MGEs

Due to the previously limited characterization of the diversity of *Nitrososphaerota*-associated viruses/MGEs, our understanding of their metabolic capacities and potential contributions to archaeal host metabolism remains largely incomplete. As with other archaeal head-tailed viruses [26,27,47], these viruses/MGEs encode proteins essential for core viral functions (DNA replication, transcription, packaging, and structural components) and proteins counteracting host defences, including restriction-modification systems and chromosome partitioning (Supplementary Dataset 6, Supplementary Figs. 3 and 4). Notably, methyltransferases were widespread (170 genomes, 140 vOTUs), occasionally coupled with endonucleases. The presence of *parB* genes in 71 genomes (36 vOTUs) suggests the likelihood of plasmid-like partitioning mechanisms in viral replication [49,50]. The newly identified *Aigarchaea*-associated rod-shaped viral family, *Byatisviridae*, was found to uniquely encode demethyl-ubiquinol-6 methylase (Supplementary Dataset 6), implying involvement of the host's ubiquinone metabolism, a metabolic potential not observed in other rod-shaped archaeal viruses[27,28,51].

Some virus-encoded proteins may serve as an auxiliary metabolic potential reservoir for the hosts. A total of 32 distinct viral-encoded protein families have no detectable homologs in host archaeal panproteomes, suggesting potential host metabolic compensation (Fig. 5 and 6a, Supplementary Figs. 8, Supplementary Dataset 7). Specific examples include 6-phosphogluconolactonase (PPP pathway) encoded by *Aigarchaea*-associated (*Wolframiiraptoraceae*) viruses/MGEs [21,23], and *sdmt* (sarcosine/dimethylglycine N-methyltransferase) in *Aigarchaea*-associated (*Caldarchaeales*) viruses/MGEs. The viral/MGE-encoded *sdmt* gene potentially aids the osmotic and oxidative stress tolerance of *Caldarchaeales* cells through betaine production (Fig. 5) [52,53]. The presence of glucose-6-phosphate isomerase (*pgi1*)

encoded by a Group 1.1b-associated viruses suggests a novel glycolytic capability of AOA (Fig. 5), previously reported only in non-AOA *Thaumarchaea* [20]. Additionally, the detection of Isopenicillin-N synthase (*pcbc*) gene indicates that viruses may potentially confer the metabolic capacity for β-lactam antibiotic production to archaea [54,55]. For Group 1.1a archaea, virus-encoded *calS13* (capsular polysaccharide biosynthesis) implies possible roles in biofilm formation, expanding archaeal capabilities in extracellular polymeric substance production [56].

Horizontal gene transfers landscape between *Nitrososphaerota*-associated viruses/MGEs and their archaeal hosts

Owing to the previously insufficient exploration of the diversity of *Nitrososphaerota*-associated viruses/MGEs, our understanding of horizontal gene transfer (HGT) dynamics between these elements and their archaeal hosts remains fragmentary. This gap is particularly pronounced for *Aigarchaea*, for which virus–host associations related to protein metabolism and HGT have not been characterized yet. We explored virus-host gene exchange, identifying 6,505 HGT events from 527 phylogenetic trees, using the approach established by Irwin et al. (Extended Data 5) [57]. Approximately 54% of gene transfers occur bidirectionally between *Nitrososphaerota* and viruses/MGEs, 45% are from *Nitrososphaerota* to viruses/MGEs, and only 0.8% are from viruses/MGEs to *Nitrososphaerota* (Fig. 6b, Supplementary Dataset 8). This highlights the significant role of viruses and MGEs in facilitating gene exchange within archaeal communities. With regard to different viral/MGE families, MGE_F003 might facilitate extensive transfer of glycosylation-related genes within Group 1.1a (Fig. 6c and Supplementary Dataset 8); MGE_F001 transferred gene for the recombination protein RecA extensively within clades of *Thaumarchaea*; *Fthagguaviridae* was primarily involved in structural-protein-associated transfers within *Aigarchaea*; *Ekchuahviridae* and *Yangangviridae* mediated glycosyltransferase and methyltransferase exchanges (Supplementary Dataset 8), respectively. This result indicates that different viral families exhibit distinct preferences in the types of

metabolic functions involved in these HGT events.

HGT events mediated by these viruses/MGEs have the potential to indirectly influence elemental cycling, prominently phosphorus (60%) and nitrogen cycles via *queC* and *queD* genes (folate biosynthesis), encoding enzymes producing orthophosphate, triphosphate, and ammonia-derived reactions (Fig. 6d and Supplementary Dataset 8) [58,59]. Cobaltochelatase transfer events suggest viral roles in cobalt metabolism, especially between Group 1.1a and *Yebviridae*. Further, HGTs involving in carbon fixation, urea cycle, TCA cycle, purine synthesis, and DNA synthesis highlight the viral contribution to carbon cycling. Additional transfers related to sulphur, manganese, halogens, and copper metabolism were identified (Fig. 6d and Supplementary Dataset 8), suggesting that archaeal viruses/MGEs may indirectly participate in extensive elemental cycling.

Discussion

These results significantly expand the current understanding of *Nitrososphaerota*-associated viruses/MGEs, broadening the scope from previously reported *Thaumarchaea*-associated viruses, and identified 263 previously unrecognized *Aigarchaea*-associated viruses for the first time, expanding the known number of *Nitrososphaerota*-associated viruses/MGEs by 10-times [30-34]. The mining of *Aigarchaea*-associated viruses/MGEs viral genomes fills a critical knowledge gap of archaeal viromes. Based on this catalogue, diverse *Nitrososphaerota*-associated viral families were identified and their host associations, genomic features, phylogenetic traits, functional potentials, and HGT landscape were described.

The phylogenetic analysis of *Nitrososphaerota*-associated viruses showed their evolutionary relationships with other viruses. In the past, only the capsids of *Lipothrixviridae* and *Tristromaviridae* were known to be composed of heterodimers[51]. Here, it was found that some *Aigarchaea*-associated rod-shaped viral capsids also consist of heterodimers (Fig. 4a, Supplementary Figs. 5 and 6), thus updating our understanding of rod-shaped archaeal viral evolution. With respect to capsid assembly

strategy, homodimers may provide symmetry and repetitiveness, thereby contributing to the maintenance of capsid stability [60,61]. Heterodimers may afford greater structural flexibility and diversity, enabling adaptation to varying environmental stresses and host defence mechanisms [62,63]. The presence of both homodimers and heterodimers in *Byatisviridae* suggests that these viruses are undergoing capsid diversification in response to environmental stress. Some *Aigarchaea*-associated viruses have close relationships with ANME-1 (*Methanobacteriota*)-associated viruses/MGEs originating from hydrothermal vents (e.g. Methanophagales virus PBV300, *Ahmunviridae*) (Supplementary Fig. 2)[28], implying viral host-switching or non-specific transfers occurred between both archaeal lineages, which is common in the evolution of viruses [64-66].

Viruses have been found to encode diverse metabolic genes that manipulate host metabolism, thereby creating favorable conditions for completing their life cycles and indirectly contributing to global biogeochemical cycles [67-70]. In archaea-associated viruses, such as those infecting methanogenic archaea, Asgard superphylum, halophilic archaea, viral-encoded metabolic genes have been identified that potentially modulate archaeal methane-oxidization, sulfur metabolism, nucleotide and amino acid metabolisms [25,71,72]. Regarding *Thaumarchaea*-associated viruses/MGEs, previous studies have discovered that viruses may manipulate host ammonia oxidation metabolism by carrying the *amoC* gene and participate in host vitamin synthesis [32,33]. Through the systematic exploration of *Nitrososphaerota*-associated viruses, we have not only expanded the functional diversity of metabolic genes encoded by *Thaumarchaea*-associated viruses/MGEs but also, more importantly, filled the gap in the functional diversity of *Aigarchaea*-associated viruses/MGEs. Metagenomic investigations of geothermal springs have revealed that *Aigarchaea* may possess the pentose phosphate pathway (PPP) but lack two core enzymes [21]: glucose-6-phosphate dehydrogenase and 6-phosphogluconolactonase. The 6-phosphogluconolactonase encoded by *Aigarchaea*-associated viruses/MGEs may compensate for the missing steps in the host's PPP (Fig. 5, Supplementary Figs. 8, Supplementary Dataset 6), thereby supporting redox balance and cellular resilience. Additionally, we have

identified, for the first time, the presence of the *sdmt* gene in viruses, suggesting that viruses may assist hosts in synthesizing betaine through the glycine betaine methylation pathway, helping them cope with high osmotic environments [73,74]. As a type of methyltransferase, *sdmt* may also function similarly to other viral-encoded methyltransferases, playing a role in immune evasion during viral-host interactions [75-77]. In this context, this represents a typical example of the pervasive exaptation of host proteins by viruses [65]. Beyond encoding genes that potentially aid host environmental adaptation, archaeal viruses also carry antibiotic resistance genes (Fig. 5, Supplementary Figs. 8, Supplementary Dataset 6). Archaea are generally unaffected by β-lactam antibiotics, and previous studies have demonstrated that archaea can produce antibiotics [78]; however, no research has yet indicated that archaea are capable of producing β-lactam antibiotics. The presence of *pcbc* genes in related viruses may help archaeal hosts gain an advantage in interspecies competition with bacteria through production of β-lactam antibiotics, thereby providing more opportunities for viruses to complete their life cycles. In summary, the significantly expanded diversity of *Nitrososphaerota*-associated viruses/MGEs revealed in this study provides prior knowledge for future in-depth research into virus-archaea interactions.

Methods

Assembly of *Nitrososphaerota*-associated CRISPR database

The *Nitrososphaerota* phylum comprises two major clades: *Thaumarchaea* and its sister clade, *Aigarchaea*. To produce a comprehensive genomic dataset, reference genomes were retrieved from GenBank and IMG/M [37,38], encompassing a total of 2,137 genomic assemblages. Specifically, 1,830 assemblages (comprising 322,784 genomic contigs) were obtained from GenBank; 307 assemblages (containing 23,395 genomic contigs) were sourced from IMG/M.

To identify CRISPR spacers within these genomes, similar to previous studies [25,26], CRASS v.0.3.8 and MinCED v.0.4.2 (CRISPR recognition tool, CRT) were employed using the following parameters to detect spacers and discard those with low-complexity and short length [79,80]: word size = 7, spacer length = 26, and a minimum of three directed repeats (DRs). This step resulted in the identification of 37,485 *Nitrososphaerota*-associated spacers, which were subsequently used for the detection of viruses/MGEs and downstream analyses.

The metagenome-derived spacers that are associated with *Nitrososphaerota* were detected to further expand the diversity of these CRISPR spacers. First, a total of 153 metagenome-derived assemblages (including 58 from the Malaspina expedition) from diverse ecosystems were retrieved and downloaded from IMG/M and ENA (European Nucleotide Archive) (Supplementary Dataset 1)[38,41]. These metagenomes were from 21 different studies, focusing on extreme natural habitats and including 92,037,046 assembled contigs. These include four studies from deep-sea, three two from sea ice, four studies from a hypersaline environment, and one study from an acidic environment. Second, the CRISPR spacers found in these contigs were identified using the same approach described above, producing 1,828,641 spacers and 1,918,441 DRs. Third, all DRs were aligned to *Nitrososphaerota* genomes using MMseqs2 with parameters [81]: word size = 7, pairwise alignment identity = 1, percentage coverage of DRs = 1, mismatches = 1. This step produced 196,551 spacers

derived from 153 metagenomes. Finally, this process generated a *Nitrososphaerota*-associated spacer database containing 234,036 (Fig. 1 and Extended Data 1).

As in previous studies [26], CRISPR-Cas cassette prediction and subtyping for archaeal contigs and genomes were conducted using CRISPRcasIdentifier v1.0.0 [82], employing a combination of Classification and Regression Trees (CART), Support Vector Machine (SVM), and Extremely Randomized Trees (ERT) algorithms for classification and regression. The analysis was performed with the following parameters: –mode ERT SVM CART, -hmm HMM1 HMM2 HMM3 HMM4 HMM5. Subtype assignment was determined through a consensus-based approach, where only subtype classifications consistently predicted by all three models (CART, SVM, and ERT) were accepted. The final subtype was selected based on the highest mean probability and the lowest standard deviation among the predictions.

Mining and filtering *Nitrososphaerota*-associated viruses/MGEs from IMG/VR (v.4)

The IMG/VR (v.4) database, contigs from 153 metagenomes, and archaeal genomes served as the primary resources for mining *Nitrososphaerota*-associated viral/MGE genomes. At first, the archaeal viral/MGE mining pipeline was validated using CRISPR spacers derived from archaeal genomes and host predictions from the IMG/VR (v.4) database and other studies to minimize false positives [31-34,36,83,84]. The 37,485 CRISPR spacers from *Nitrososphaerota* genomes were aligned to IMG/VR (v.4) database using MMseqs2 with the parameters: word size = 7, pairwise alignment identity = 0.9, percentage coverage of spacers = 0.9, mismatches = 1, resulting in 368 targeted. Among these results, the hosts of 46 viruses/MGEs were also predicted to be *Nitrososphaerota* in IMG/VR (v.4). Due to some viral contigs being assigned to non-*Nitrososphaerota* hosts in IMG/VR (v.4) metadata, the genome-content similarity between viruses/MGEs and hosts was applied to reduce the false positives: i) the open reading frames (ORFs) of these viral contigs were predicted by Prodigal v.2.6.3 and aligned to the NR database (2024.05.20) with 1e-5 as the E-value cut-off [85]; ii) the

Nitrososphaerota-related hits were counted and normalized based on viral contig lengths; iii) for each genome, it was observed that at least 20% of ORFs per 1-kb genome could be matched to *Nitrososphaerota* (Supplementary Dataset 2); iv) further validation was conducted to assess the applicability of this screening criterion to previously reported *Nitrososphaerota*-associated viruses/MGEs:

1) the 13 identified *Nitrososphaerota*-associated viral/MGE genomes were retrieved and downloaded from GenBank;
2) the 11 representative genomes were generated from 13 *Nitrososphaerota*-associated viral genomes through MMseqs2 with parameters [81]: pairwise alignment identity = 0.9, pairwise alignment coverage for shorter sequences = 0.9;
3) the 9 out of 11 viral genomes derived from GenBank meet this criterion. Two isolated spindle-shaped viruses (NSV4 and NSV5) failed to pass the test [35]. This indicates that the approach is also applicable to most previously reported *Nitrososphaerota*-associated viral/MGE genomes.

After validating the filtering approach, it was used as a cut-off for viral-host genome-content similarity to further filter the results obtained from spacer-protospacer matching. By applying this approach, 171 *Nitrososphaerota*-associated viruses/MGEs were filtered from 368 spacer-targeted viral/MGE candidates. Similar to previous studies [27,47], all relevant archaeal CRISPR-Cas systems were evaluated using CRISPRCasFinder with default parameters [86]. Among the identified CRISPR arrays, 91.2% (n = 301) exhibited an evidence level of 4, characterized by a repeat conservation index exceeding 70% and spacer identity below 8% (Supplementary Dataset 2). The remaining 29 viruses/MGEs, corresponding to evidence levels below 4 for CRISPR arrays, were retained for two reasons: 1) they passed genome-content similarity filtering; 2) in genome-wide analyses, they clustered with other viruses/MGEs possessing CRISPR arrays with an evidence level of 4.

In addition, IMG/VR (v.4) contains 1,457 viruses/MGEs that were predicted to be associated with *Nitrososphaerota*, although these contigs were not retrieved through spacer-protospacer matching. The 1,457 viruses/MGEs predicted to be associated with *Nitrososphaerota* in IMG/VR (v.4) were tested. First, 699 representative contigs were

clustered using MMSeq2 with parameters: pairwise alignment identity = 0.9, pairwise alignment coverage for shorter sequences = 0.9. Second, these contigs were tested by genome-content similarity calculation, resulting in 441 *Nitrososphaerota*-associated viruses/MGEs. Finally, 612 *Nitrososphaerota*-associated viruses/MGEs were identified and filtered from IMG/VR (v.4).The completeness of these viral/MGEs were evaluate using CheckV [87].

Expanded mining of *Nitrososphaerota*-associated viruses/MGEs from 153 metagenomes

The *Nitrososphaerota*-associated CRISPR spacer database was aligned to from 153 metagenomes using MMseqs2 with the same approach and parameters [81]. Contigs containing DRs and those shorter than 2-kb were removed before performing alignment. This resulted in 1,951 contigs matched by *Nitrososphaerota*-associated CRISPR spacers. To remove duplications, 1,951 contigs were aligned to previously identified *Nitrosospaherota*-associated viruses/MGEs set using MMseqs2 with same parameters same parameters [81]. This step generated 1,895 contigs, including 1,396 representative contigs clustered using the same approach described before. To recall as many as possible viruses/MGEs, these contigs were used to filter *Nitrososphaerota*-associated viruses/MGEs using three multiple viral identification pipelines, including geNomad, VirSorter2, and DeepVirFinder[88-90]. Multiple pipelines were employed for virus identification based on the following considerations: 1) geNomad demonstrates superior overall classification performance compared to other pipelines [89]; 2) VirSorter2 (VS2) is recommended when integrating multiple pipelines [91]; 3) DeepVirFinder (DVF), utilizing an alignment-free approach, is well-suited for identifying unique viral sequences [88]; 4) the identification results from each pipeline complement one another, enhancing recall while mitigating cellular contamination in the dataset[92-94]. For each pipeline, precision was prioritized over sensitivity; therefore, stringent virus detection strategies were adopted to minimize potential false positives: 1) geNomad was used to detect viral/MGE contigs using conserved mode ("--

conservative"), resulting in a more strict set of viruses with high confidence; 2) VS2 was used to detect viral contigs under high-confidence mode ('--high-confidence-only'), retaining results with classification score ≥ 0.9, and classification score ≤ 0.7 with viral hallmark ≥ 1; 3) The DVF used to detect viral contig with parameters: score ≥ 0.9 and p-value ≤ 0.01. To validate this hybrid viral detection pipeline, 20% of sequences were randomly selected from previously identified *Nitrososphaerota*-associated viruses/MGE and were used as a test set to evaluate the efficiency of the approach. Based on the test, this viral/MGE identification approach achieved a 74% recruitment rate, which was considered acceptable for further viral/MGE identification (Supplementary Dataset 9).

Using the validated hybrid viral identification approach, 419 viral/MGE contigs were identified from 1,396 metagenome-derived contigs. These contigs had already been targeted by *Nitrososphaerota*-associated CRISPR spacers in previous steps. The viral-host genome-content similarity was used to filter these contigs as described before, resulting 108 *Nitrososphaerota*-associated viral/MGEs. Relevant archaeal CRISPR-Cas systems of these contigs were evaluated using CRISPRCasFinder with default parameters [86]. The completeness of these viral/MGEs was evaluated using CheckV [87].

Detection of viruses/MGEs integrated in *Nitrososphaerota* genomes

Proviruses integrated into *Nitrososphaerota* genomes were identified using the "find-proviruses" program in geNomad [89]. Integrases were first detected with the MMseqs2 plugin in geNomad [81], applying an E-value cut-off of 1e-5. Additional parameters were set as follows: a minimum gene-level score of 0.4 for flagging proviral genes using the conditional random field model; a minimum total virus marker score of 12 for proviruses lacking integrases or not located at scaffold edges; a minimum total virus marker score of 8 for proviruses encoding integrases or located at scaffold edges; a maximum allowed distance of 10-kb between provirus boundaries and integrases for boundary extension; and a maximum allowed distance of 5-kb between provirus

boundaries and tRNAs for boundary extension. This process resulted in the identification of 30 *Nitrososphaerota*-associated proviruses.

To remove redundancy, the 30 proviral/MGE genomes performed self-alignment and were compared against previously identified *Nitrososphaerota*-associated viral/MGE genomes using the same approach and parameters described above. This filtering step yielded 24 non-redundant *Nitrososphaerota*-associated proviruses (Supplementary Dataset 2).

Host assignments of *Nitrososphaerota*-associated viruses/MGEs

The host assignments of *Nitrososphaerota*-associated viruses/MGEs were primarily based on: 1) CRISPR-spacers-associated host genomes; 2) information provided by previous studies; 3) genome-content similarity between viruses/MGEs and hosts. For *Nitrososphaerota*-associated viruses/MGEs that match CRISPR spacers from multiple distinct hosts, similar with previous study [27], the host of a virus/MGE targeted by CRISPR spacers was assigned based on the taxonomy of the genome from which the spacer with the highest identity originated. For *Nitrososphaerota*-associated viruses/MGEs that were identified based on oligonucleotide pattern similarity, the most possible host genomes associated with these viruses/MGEs were determined through alignment against *Nitrososphaerota* genomes using MMseqs2 with the parameters: 1e-5 as the E-value, iteratively searched two times [81]. A host genome was linked to a viral/MGE genome if they have the most shared ORFs. This approach was also applied in IMG/VR (v.4) [83,95]. For *Nitrososphaerota*-associated viruses/MGEs reported in previous studies, some host information was retrieved from related literature. For those viruses/MGEs with uncertain host information [33], the related viral ORFs were subject to MMseqs2 against all *Nitrososphaerota* genomes with the same approach. The 35 Marthaviruses derived from López-Pérez et al. study were excluded from this study for two reasons [32]: 1) the related data (PRJNA484324) are still embargoed when an attempt was made to access these genomes; 2) although the proposed taxa 'Marthaviruses' has been documented in NCBI, the host information of

these viruses is labelled as bacteria. The archaeal host phylogenetic tree was calculated based on 46 concatenated ribosomal proteins by IQ-Tree2 (Supplementary Note 2)[96]. The archaeal lineages were determined based on their phylogenetic placements and previous studies [2,8,20,21,23,46].

Comparative genomics and family-/species-level demarcation of *Nitrososphaerota*-associated viruses/MGEs

Similar to previous studies [25,27,29], new families and vOTUs were assigned based on the genome-content similarity network and the genome-wide proteomic tree. Using DNA virus sequences longer than 2-kb from NCBI RefSeq v.223 as reference genomes, vConTACT2 was applied (with default parameters) to compute the genome-content-based similarity network of those *Nitrososphaerota*-associated viruses/MGEs. *Nitrososphaerota*-associated viral/MGE genome pairs that shared at least 20% protein clusters and encoded at least six genes were filtered from genome-content-based similarity network, generating a subset of network, which was used to guide taxonomic assignments. In the network, viruses not related to *Nitrososphaerota*-associated viruses/MGEs were discarded to achieve a clear layout. Genome-wide proteomic tree of those genomes included in the subset of network were calculated by ViPTree v.4. The viruses/MGEs were assigned to the same family if they were represented by a cohesive group in the genome-content-based similarity network and a monophyletic group in genome-wide proteomic tree (with threshold 0.05) [27]. The network and tree were visualized by Cytoscape v.3.10.2 and iToL v.6, respectively [97,98]. The nomenclature of the proposed viral family was referred to Cthulhu mythos (Supplementary Note 3).

Annotation of viral/MGEs genomes

Viral/MGE species (or vOTUs) were defined using 95% nucleotide identity threshold over 90% of the full length using MMseqs2. The family-level taxonomic assignments of *Nitrososphaerota*-associated viruses/MGEs were assigned using 1) VITAP with

custom database: this database includes all viral reference genomes collected in ICTV VMR-MSL39, high-quality vOTUs with genus-level taxonomy collected in IMG/VR (v.4), and all viral genomes collected in NCBI RefSeq v.223 [83,99,100]; 2) CAT with NCBI RefSeq v.223 database [100,101]; 3) geNomad with default parameters [89]. The CAT only assigned class/family level units for 10 contigs; the geNomad and VITAP assigned class/family level units for 683 and 676 contigs, respectively. Hence, the taxonomic assignments were mainly based on geNomad and VITAP. If geNomad or VITAP showed consistency at higher taxonomic levels (Phylum or Class), but one of the pipelines produced annotations at lower taxonomic levels, then the results with the lower taxonomic level annotations were used. The results with discrepancies between VITAP and geNomad were compared against the NCBI RefSeq v.223 using MMseqs2 with default parameters and subsequently subjected to manual verification. The consensus VCs generated from vConTACT2 were used to expanded and furtherly curated the viral taxonomic assignment [102]. The members of each proposed family were furtherly expanded based on consensus vOTU.

ORFs encoded in *Nitrososphaerota*-associated viruses/MGEs were predicted by Prodigal v.2.6.3 with metagenome mode[85], and annotated using the KEGG Ortholog database (January, 2024), Pfam-A v.35, COG2020, Swiss-Prot database (January, 2024), and VOG223 database [103-105].The pan-proteomes of related host lineages were aligned to KEGG Ortholog database with the same parameters. The metabolic pathways involved in viruses and their hosts were visualized using iPath 3 [106].

Determining of *Nitrososphaerota*-associated viral/MGE horizontal gene transfer events

HGT events between *Nitrososphaerota*-associated viruses (vOTUs) and their archaeal hosts were identified following an established pipeline [57]. After removal of proviral sequences, viral and archaeal sequences were aligned, and shared protein families (PFs) were identified. Potential contamination was rigorously screened based on lineage, GC content, scaffold length, and genomic context, retaining only robust

alignments (Supplementary Note 4 and Supplementary Dataset 10).

Filtered protein families underwent phylogenetic analysis using MUSCLE5 alignment, trimAl v.1.4 trimming, and IQ-TREE2 tree construction with the LG+F+R5 model. Trees were rooted via minimal ancestral deviation, and potential HGT events were classified based on the phylogenetic relationships between viral and archaeal sequences (Extended Data 5). Transfer direction labels ("v-from-a", "v-to-a", "v-?-a" and "non-HGT") were assigned following predefined criteria considering sister and cousin clade compositions. Iterative screening was performed with temporary reassignment of labels to ensure accurate HGT detection across multiple rounds of analysis (Supplementary Note 4).

Phylogenetic inference of *Nitrososphaerota*-associated viruses/MGEs

To investigate the evolutionary relationships among rod-shaped and head-tailed viruses associated with *Nitrososphaerota* within the virosphere context, phylogenetic analyses were performed based on key viral proteins, including the SIRV2-like major capsid protein (MCP), HK97-like MCP, and terminase proteins. For rod-shaped viruses, MCP sequences from 33viruses/MGEs were aligned with 34 viral RefSeqsv.223 from the *Adnaviria* realm using MMSeqs2 (E-value ≤1e-5). Notably, some viruses encoded adjacent duplicate MCP subunits, including Metallosphaera rod-shaped virus 1 and Saccharolobussolfataricus rod-shaped virus 1, along with five uncultured viral genomes (UViGs). Considering the fundamental rod-shaped viral MCP structural unit as either homodimers or heterodimers, phylogenetic trees were inferred based on concatenated MSAs generated using MUSCLE5 (ensemble mode). The DUF732 domain-containing protein (PF05305), identified via FoldSeek structural alignment as the closest homolog [107], served as an outgroup for rooting the rod-shaped viral MCP phylogeny. HK97-MCP sequences exhibited significant diversity, encompassing multiple families from Pfam-A v.35 and VOG223 [104,108]; thus, three dominant domains (PF05065, PF07068, and VOG01305) were selected and aligned against *Duplodnaviria* proteins from RefSeqs v.223. Appropriate outgroups were

chosen based on structural and domain-based alignments (e.g., PF06673 and PF00130). Similarly, terminase proteins were analyzed by selecting the top three domains (K06909, VOG00195, and PF03237) and employing P-loop containing nucleoside triphosphate hydrolases (SSF52540) as an outgroup. All MSAs were refined using trimAl v.1.4, and trees were inferred by ML with IQ-Tree2 and BI using MrBayes [96,109], supported by 1,000 bootstrap (ML) and convergence diagnostics (BI) (Extended Data 3 and 4). The topological comparisons between BI and ML were performed using Phylo.io [110].

Structure analysis of *Nitrososphaerota*-associated viruses/MGEs

To further assess the novelty of capsid proteins from *Nitrososphaerota*-associated viruses/MGEs, the HK97-like and SIRV2-like capsid proteins identified in this study were used to perform structural prediction. A total of 5,153 head-tailed viral genomes from RefSeqs v.223 and 44 rod-shaped archaeal viral genomes from GenBank (length ≥ 2,000 bp) were retrieved from GenBank (as of November 31, 2024). ORFs in these genomes were predicted using Prodigal v.2.6.3 [85]. HK97-like and SIRV2-like capsid proteins were identified using HMM-based searches against specific HMM profiles (PF12193, PF05065, VOG01305, and PF07068) with an E-value cut-off of 1e-5 [111], resulting in 1,844 HK97-like capsids and 28 SIRV2-like capsids. Subsequently, the HK97-like capsid proteins were clustered using MMSeqs2 with thresholds of 0.9 alignment coverage and identity [81], yielding 468 representative sequences. Structures of the HK97-like and SIRV2-like capsid proteins were predicted using AlphaFold2 implemented in ColabFold (Extended Data 6 and 7) [112], and the models with the highest predicted local distance difference test (pLDDT) scores were selected as best-predicted models. Structural alignments were conducted through all-to-all comparisons using FoldSeek with an identity cut-off of 0.5 [107], and reciprocal hits were used to construct a structural similarity network. The structural similarity network was built based on bitscores (top 10% hits were selected) calculated from FoldSeek by Cytoscape[97,107].

Data availability

All nucleotide and protein sequences of viral or mobile genetic elements, spacers from the CRISPR database, phylogenetic trees, structural prediction files are available in FigShare (https://figshare.com/s/a2e94d686028b674a905). The related metagenomes used in this study can be accessed via the URL provided in Supplementary Dataset 1 through IMG/M. Related genomes from IMG/VR database is available at https://img.jgi.doe.gov/cgi-bin/vr/main.cgi. Other information and metadata are provided as Supplementary Datasets.

Code availability

No custom code was generated and used.

Reference

1. Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., Chee, G.J., Hattori, M., Kanai, A., Atomi, H., et al. (2011). Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res *39*, 3204-3223. 10.1093/nar/gkq1228.
2. Spang, A., Caceres, E.F., and Ettema, T.J.G. (2017). Genomic exploration of the diversity, ecology, and evolution of the archaeal domain of life. Science *357*. 10.1126/science.aaf3883.
3. Alfreider, A., Grimus, V., Luger, M., Ekblad, A., Salcher, M.M., and Summerer, M. (2018). Autotrophic carbon fixation strategies used by nitrifying prokaryotes in freshwater lakes. FEMS Microbiol Ecol *94*. 10.1093/femsec/fiy163.
4. Klotz, F., Kitzinger, K., Ngugi, D.K., Busing, P., Littmann, S., Kuypers, M.M.M., Schink, B., and Pester, M. (2022). Quantification of archaea-driven freshwater nitrification from single cell to ecosystem levels. ISME J *16*, 1647-1656. 10.1038/s41396-022-01216-9.
5. Li, Y., Chen, J., Lin, Y., Zhong, C., Jing, H., and Liu, H. (2024). Thaumarchaeota from deep-sea methane seeps provide novel insights into their evolutionary history and ecological implications. Microbiome *12*, 197. 10.1186/s40168-024-01912-y.
6. Brochier-Armanet, C., Boussau, B., Gribaldo, S., and Forterre, P. (2008). Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol *6*, 245-252. 10.1038/nrmicro1852.
7. Pester, M., Schleper, C., and Wagner, M. (2011). The Thaumarchaeota: an emerging view of their phylogeny and ecophysiology. Curr Opin Microbiol *14*, 300-306. 10.1016/j.mib.2011.04.007.
8. Zhong, H., Lehtovirta-Morley, L., Liu, J., Zheng, Y., Lin, H., Song, D., Todd, J.D., Tian, J., and Zhang, X.H. (2020). Novel insights into the Thaumarchaeota in the deepest oceans: their metabolism and potential adaptation mechanisms. Microbiome *8*, 78. 10.1186/s40168-020-00849-2.
9. Kuypers, M.M.M., Marchant, H.K., and Kartal, B. (2018). The microbial nitrogen-cycling network. Nat Rev Microbiol *16*, 263-276. 10.1038/nrmicro.2018.9.
10. Prosser, J.I., and Nicol, G.W. (2012). Archaeal and bacterial ammonia-oxidisers in soil: the quest for niche specialisation and differentiation. Trends

Microbiol *20*, 523-531. 10.1016/j.tim.2012.08.001.

11. Martens-Habbena, W., Berube, P.M., Urakawa, H., de la Torre, J.R., and Stahl, D.A. (2009). Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. Nature *461*, 976-979. 10.1038/nature08465.

12. Stahl, D.A., and de la Torre, J.R. (2012). Physiology and diversity of ammonia-oxidizing archaea. Annu Rev Microbiol *66*, 83-101. 10.1146/annurev-micro-092611-150128.

13. Walker, C.B., de la Torre, J.R., Klotz, M.G., Urakawa, H., Pinel, N., Arp, D.J., Brochier-Armanet, C., Chain, P.S., Chan, P.P., Gollabgir, A., et al. (2010). Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. Proc Natl Acad Sci U S A *107*, 8818-8823. 10.1073/pnas.0913533107.

14. Könneke, M., Schubert, D.M., Brown, P.C., Hügler, M., Standfest, S., Schwander, T., Schada von Borzyskowski, L., Erb, T.J., Stahl, D.A., and Berg, I.A. (2014). Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO2 fixation. Proceedings of the National Academy of Sciences *111*, 8239-8244.

15. Ren, M., Feng, X., Huang, Y., Wang, H., Hu, Z., Clingenpeel, S., Swan, B.K., Fonseca, M.M., Posada, D., Stepanauskas, R., et al. (2019). Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. ISME J *13*, 2150-2161. 10.1038/s41396-019-0418-8.

16. Francis, C.A., Roberts, K.J., Beman, J.M., Santoro, A.E., and Oakley, B.B. (2005). Ubiquity and diversity of ammonia-oxidizing archaea in water columns and sediments of the ocean. Proc Natl Acad Sci U S A *102*, 14683-14688. 10.1073/pnas.0506625102.

17. Nelkner, J., Huang, L., Lin, T.W., Schulz, A., Osterholz, B., Henke, C., Blom, J., Puhler, A., Sczyrba, A., and Schluter, A. (2023). Abundance, classification and genetic potential of Thaumarchaeota in metagenomes of European agricultural soils: a meta-analysis. Environ Microbiome *18*, 26. 10.1186/s40793-023-00479-9.

18. Biggs-Weber, E., Aigle, A., Prosser, J.I., and Gubry-Rangin, C. (2020). Oxygen preference of deeply-rooted mesophilic thaumarchaeota in forest soil. Soil Biology and Biochemistry *148*, 107848. https://doi.org/10.1016/j.soilbio.2020.107848.

19. Aylward, F.O., and Santoro, A.E. (2020). Heterotrophic Thaumarchaea with Small Genomes Are Widespread in the Dark Ocean. mSystems *5*. 10.1128/mSystems.00415-20.

20. Zhang, R.Y., Wang, Y.R., Liu, R.L., Rhee, S.K., Zhao, G.P., and Quan, Z.X. (2024). Metagenomic characterization of a novel non-ammonia-oxidizing

Thaumarchaeota from hadal sediment. Microbiome *12*, 7. 10.1186/s40168-023-01728-2.

21. Hua, Z.S., Qu, Y.N., Zhu, Q., Zhou, E.M., Qi, Y.L., Yin, Y.R., Rao, Y.Z., Tian, Y., Li, Y.X., Liu, L., et al. (2018). Genomic inference of the metabolism and evolution of the archaeal phylum Aigarchaeota. Nat Commun *9*, 2832. 10.1038/s41467-018-05284-4.
22. Beam, J.P., Jay, Z.J., Schmid, M.C., Rusch, D.B., Romine, M.F., Jennings Rde, M., Kozubal, M.A., Tringe, S.G., Wagner, M., and Inskeep, W.P. (2016). Ecophysiology of an uncultivated lineage of Aigarchaeota from an oxic, hot spring filamentous 'streamer' community. ISME J *10*, 210-224. 10.1038/ismej.2015.83.
23. Buessecker, S., Palmer, M., Lai, D., Dimapilis, J., Mayali, X., Mosier, D., Jiao, J.Y., Colman, D.R., Keller, L.M., St John, E., et al. (2022). An essential role for tungsten in the ecology and evolution of a previously uncultivated lineage of anaerobic, thermophilic Archaea. Nat Commun *13*, 3773. 10.1038/s41467-022-31452-8.
24. Prangishvili, D., Bamford, D.H., Forterre, P., Iranzo, J., Koonin, E.V., and Krupovic, M. (2017). The enigmatic archaeal virosphere. Nat Rev Microbiol *15*, 724-739. 10.1038/nrmicro.2017.125.
25. Medvedeva, S., Sun, J., Yutin, N., Koonin, E.V., Nunoura, T., Rinke, C., and Krupovic, M. (2022). Three families of Asgard archaeal viruses identified in metagenome-assembled genomes. Nat Microbiol *7*, 962-973. 10.1038/s41564-022-01144-6.
26. Rambo, I.M., Langwig, M.V., Leao, P., De Anda, V., and Baker, B.J. (2022). Genomes of six viruses that infect Asgard archaea from deep-sea sediments. Nat Microbiol *7*, 953-961. 10.1038/s41564-022-01150-8.
27. Medvedeva, S., Borrel, G., Krupovic, M., and Gribaldo, S. (2023). A compendium of viruses from methanogenic archaea reveals their diversity and adaptations to the gut environment. Nat Microbiol *8*, 2170-2182. 10.1038/s41564-023-01485-w.
28. Laso-Perez, R., Wu, F., Cremiere, A., Speth, D.R., Magyar, J.S., Zhao, K., Krupovic, M., and Orphan, V.J. (2023). Evolutionary diversification of methanotrophic ANME-1 archaea and their expansive virome. Nat Microbiol. 10.1038/s41564-022-01297-4.
29. Sakai, H.D., Nur, N., Kato, S., Yuki, M., Shimizu, M., Itoh, T., Ohkuma, M., Suwanto, A., and Kurosawa, N. (2022). Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses. Proc Natl Acad Sci U S A *119*. 10.1073/pnas.2115449119.
30. Krupovic, M., Spang, A., Gribaldo, S., Forterre, P., and Schleper, C. (2011). A

thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. Biochem Soc Trans *39*, 82-88. 10.1042/BST0390082.

31. Krupovic, M., Makarova, K.S., Wolf, Y.I., Medvedeva, S., Prangishvili, D., Forterre, P., and Koonin, E.V. (2019). Integrated mobile genetic elements in Thaumarchaeota. Environ Microbiol *21*, 2056-2078. 10.1111/1462-2920.14564.

32. Lopez-Perez, M., Haro-Moreno, J.M., de la Torre, J.R., and Rodriguez-Valera, F. (2019). Novel Caudovirales associated with Marine Group I Thaumarchaeota assembled from metagenomes. Environ Microbiol *21*, 1980-1988. 10.1111/1462-2920.14462.

33. Ahlgren, N.A., Fuchsman, C.A., Rocap, G., and Fuhrman, J.A. (2019). Discovery of several novel, widespread, and ecologically distinct marine Thaumarchaeota viruses that encode amoC nitrification genes. ISME J *13*, 618-631. 10.1038/s41396-018-0289-4.

34. Zhou, Y., Zhou, L., Yan, S., Chen, L., Krupovic, M., and Wang, Y. (2023). Diverse viruses of marine archaea discovered using metagenomics. Environ Microbiol *25*, 367-382. 10.1111/1462-2920.16287.

35. Kim, J.G., Kim, S.J., Cvirkaite-Krupovic, V., Yu, W.J., Gwak, J.H., Lopez-Perez, M., Rodriguez-Valera, F., Krupovic, M., Cho, J.C., and Rhee, S.K. (2019). Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. Proc Natl Acad Sci U S A *116*, 15645-15650. 10.1073/pnas.1905682116.

36. Liu, Y., Brandt, D., Ishino, S., Ishino, Y., Koonin, E.V., Kalinowski, J., Krupovic, M., and Prangishvili, D. (2019). New archaeal viruses discovered by metagenomic analysis of viral communities in enrichment cultures. Environ Microbiol *21*, 2002-2014. 10.1111/1462-2920.14479.

37. Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.M., Huntemann, M., et al. (2021). A genomic catalog of Earth's microbiomes. Nat Biotechnol *39*, 499-509. 10.1038/s41587-020-0718-6.

38. Chen, I.A., Chu, K., Palaniappan, K., Ratner, A., Huang, J., Huntemann, M., Hajek, P., Ritter, S., Varghese, N., Seshadri, R., et al. (2021). The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res *49*, D751-D763. 10.1093/nar/gkaa939.

39. Sharma, S., Lee, M., Reinmann, C.S., Pumneo, J., Cutright, T.J., and Senko, J.M. (2020). Impact of acid mine drainage chemistry and microbiology on the development of efficient Fe removal activities. Chemosphere *249*, 126117. 10.1016/j.chemosphere.2020.126117.

40. Zhang, X., Xu, W., Liu, Y., Cai, M., Luo, Z., and Li, M. (2018). Metagenomics

Reveals Microbial Diversity and Metabolic Potentials of Seawater and Surface Sediment From a Hadal Biosphere at the Yap Trench. Front Microbiol *9*, 2402. 10.3389/fmicb.2018.02402.

41. Acinas, S.G., Sanchez, P., Salazar, G., Cornejo-Castillo, F.M., Sebastian, M., Logares, R., Royo-Llonch, M., Paoli, L., Sunagawa, S., Hingamp, P., et al. (2021). Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol *4*, 604. 10.1038/s42003-021-02112-2.
42. Karner, M.B., DeLong, E.F., and Karl, D.M. (2001). Archaeal dominance in the mesopelagic zone of the Pacific Ocean. Nature *409*, 507-510. 10.1038/35054051.
43. Schattenhofer, M., Fuchs, B.M., Amann, R., Zubkov, M.V., Tarran, G.A., and Pernthaler, J. (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. Environ Microbiol *11*, 2078-2093. 10.1111/j.1462-2920.2009.01929.x.
44. Pester, M., Rattei, T., Flechl, S., Grongroft, A., Richter, A., Overmann, J., Reinhold-Hurek, B., Loy, A., and Wagner, M. (2012). amoA-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoA genes from soils of four different geographic regions. Environ Microbiol *14*, 525-539. 10.1111/j.1462-2920.2011.02666.x.
45. Vetriani, C., Jannasch, H.W., MacGregor, B.J., Stahl, D.A., and Reysenbach, A.L. (1999). Population structure and phylogenetic characterization of marine benthic Archaea in deep-sea sediments. Appl Environ Microbiol *65*, 4375-4384. 10.1128/AEM.65.10.4375-4384.1999.
46. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol *2*, 1533-1542. 10.1038/s41564-017-0012-7.
47. Wu, Z., Liu, S., and Ni, J. (2024). Metagenomic characterization of viruses and mobile genetic elements associated with the DPANN archaeal superphylum. Nat Microbiol. 10.1038/s41564-024-01839-y.
48. Liu, Y., Demina, T.A., Roux, S., Aiewsakun, P., Kazlauskas, D., Simmonds, P., Prangishvili, D., Oksanen, H.M., and Krupovic, M. (2021). Diversity, taxonomy, and evolution of archaeal viruses of the class Caudoviricetes. PLoS Biol *19*, e3001442. 10.1371/journal.pbio.3001442.
49. Breier, A.M., and Grossman, A.D. (2007). Whole-genome analysis of the chromosome partitioning and sporulation protein Spo0J (ParB) reveals spreading and origin-distal sites on the Bacillus subtilis chromosome. Mol Microbiol *64*, 703-718. 10.1111/j.1365-2958.2007.05690.x.

50. Chu, C.H., Wu, C.T., Lin, M.G., Yen, C.Y., Wu, Y.Z., Hsiao, C.D., and Sun, Y.J. (2024). Insights into the molecular mechanism of ParABS system in chromosome partition by HpParA and HpParB. Nucleic Acids Res *52*, 7321-7336. 10.1093/nar/gkae450.
51. Krupovic, M., Kuhn, J.H., Wang, F., Baquero, D.P., Dolja, V.V., Egelman, E.H., Prangishvili, D., and Koonin, E.V. (2021). Adnaviria: a New Realm for Archaeal Filamentous Viruses with Linear A-Form Double-Stranded DNA Genomes. J Virol *95*, e0067321. 10.1128/JVI.00673-21.
52. Nyyssola, A., Kerovuo, J., Kaukinen, P., von Weymarn, N., and Reinikainen, T. (2000). Extreme halophiles synthesize betaine from glycine by methylation. J Biol Chem *275*, 22196-22201. 10.1074/jbc.M910111199.
53. Lee, Y.R., Lin, T.S., Lai, S.J., Liu, M.S., Lai, M.C., and Chan, N.L. (2016). Structural Analysis of Glycine Sarcosine N-methyltransferase from Methanohalophilus portucalensis Reveals Mechanistic Insights into the Regulation of Methyltransferase Activity. Sci Rep *6*, 38071. 10.1038/srep38071.
54. Chen, S., Dai, Y., Ke, J., Luo, Y., Wang, C., Hao, Y., Zhang, A., Han, J., and Xiang, H. (2024). Halocin H4 is activated through cleavage by halolysin HlyR4. Appl Environ Microbiol *90*, e0228423. 10.1128/aem.02284-23.
55. Torres, M.D.T., Wan, F., and de la Fuente-Nunez, C. (2024). Discovery of antibiotics in the archaeome using deep learning. bioRxiv. 10.1101/2024.11.15.623859.
56. Wang, P., Li, M., Dong, L., Zhang, C., and Xie, W. (2022). Comparative Genomics of Thaumarchaeota From Deep-Sea Sponges Reveal Their Niche Adaptation. Front Microbiol *13*, 869834. 10.3389/fmicb.2022.869834.
57. Irwin, N.A.T., Pittis, A.A., Richards, T.A., and Keeling, P.J. (2022). Systematic evaluation of horizontal gene transfer between eukaryotes and viruses. Nat Microbiol *7*, 327-336. 10.1038/s41564-021-01026-3.
58. Thiaville, J.J., Kellner, S.M., Yuan, Y., Hutinet, G., Thiaville, P.C., Jumpathong, W., Mohapatra, S., Brochier-Armanet, C., Letarov, A.V., Hillebrand, R., et al. (2016). Novel genomic island modifies DNA with 7-deazaguanine derivatives. Proc Natl Acad Sci U S A *113*, E1452-1459. 10.1073/pnas.1518570113.
59. Gaur, R., and Varshney, U. (2005). Genetic analysis identifies a function for the queC (ybaX) gene product at an initial step in the queuosine biosynthetic pathway in Escherichia coli. J Bacteriol *187*, 6893-6901. 10.1128/JB.187.20.6893-6901.2005.
60. Tubiana, T., Boulard, Y., and Bressanelli, S. (2017). Dynamics and asymmetry in the dimer of the norovirus major capsid protein. PLoS One *12*, e0182056.

10.1371/journal.pone.0182056.

61. Hadden, J.A., Perilla, J.R., Schlicksup, C.J., Venkatakrishnan, B., Zlotnick, A., and Schulten, K. (2018). All-atom molecular dynamics of the HBV capsid reveals insights into biological function and cryo-EM resolution limits. Elife *7*. 10.7554/eLife.32478.

62. Mueller, K.L. (2016). Structural insights into capsid flexibility. Science *354*, 1387-1389. 10.1126/science.354.6318.1387-p.

63. Budean, D., Almeida-Hernández, Y., Pandey, J., Pérez, J.M., Sánchez García, E., and Haglund, E. (2024). Exploring Chemokine Homodimer Stability: Structural Insights into CXC and CC Interfaces. Biophysica *4*, 561-572.

64. Krupovic, M., Dolja, V.V., and Koonin, E.V. (2019). Origin of viruses: primordial replicators recruiting capsids from hosts. Nat Rev Microbiol *17*, 449-458. 10.1038/s41579-019-0205-6.

65. Koonin, E.V., Dolja, V.V., and Krupovic, M. (2022). The logic of virus evolution. Cell Host Microbe *30*, 917-929. 10.1016/j.chom.2022.06.008.

66. Hwang, Y., Roux, S., Coclet, C., Krause, S.J.E., and Girguis, P.R. (2023). Viruses interact with hosts that span distantly related microbial domains in dense hydrothermal mats. Nat Microbiol *8*, 946-957. 10.1038/s41564-023-01347-5.

67. Hurwitz, B.L., and U'Ren, J.M. (2016). Viral metabolic reprogramming in marine ecosystems. Curr Opin Microbiol *31*, 161-168. 10.1016/j.mib.2016.04.002.

68. Tian, F., Wainaina, J.M., Howard-Varona, C., Dominguez-Huerta, G., Bolduc, B., Gazitua, M.C., Smith, G., Gittrich, M.R., Zablocki, O., Cronin, D.R., et al. (2024). Prokaryotic-virus-encoded auxiliary metabolic genes throughout the global oceans. Microbiome *12*, 159. 10.1186/s40168-024-01876-z.

69. Kieft, K., Zhou, Z., Anderson, R.E., Buchan, A., Campbell, B.J., Hallam, S.J., Hess, M., Sullivan, M.B., Walsh, D.A., Roux, S., and Anantharaman, K. (2021). Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. Nat Commun *12*, 3503. 10.1038/s41467-021-23698-5.

70. Huang, X., Zhou, Z., Liu, H., Li, Y., Ge, T., Tang, X., He, Y., Ma, B., Xu, J., Anantharaman, K., and Li, Y. (2024). Soil nutrient conditions alter viral lifestyle strategy and potential function in phosphorous and nitrogen metabolisms. Soil Biology and Biochemistry *189*, 109279. https://doi.org/10.1016/j.soilbio.2023.109279.

71. Wang, L., Wang, Y., Huang, X., Ma, R., Li, J., Wang, F., Jiao, N., and Zhang, R. (2022). Potential metabolic and genetic interaction among viruses, methanogen and methanotrophic archaea, and their syntrophic partners. ISME Communications *2*, 50. 10.1038/s43705-022-00135-2.

72. Zhou, Y., Gutiérrez-Preciado, A., Liu, Y., Moreira, D., Yakimov, M.M., López-García, P., and Krupovic, M. (2025). Nested parasitism in hypersaline environments: viruses and virus satellites of haloarchaea and their nanosized cellular symbionts. bioRxiv, 2025.2002.2015.638363. 10.1101/2025.02.15.638363.
73. D'Souza-Ault, M.R., Smith, L.T., and Smith, G.M. (1993). Roles of N-acetylglutaminylglutamine amide and glycine betaine in adaptation of Pseudomonas aeruginosa to osmotic stress. Appl Environ Microbiol *59*, 473-478. 10.1128/aem.59.2.473-478.1993.
74. Arumugam, M.K., Paal, M.C., Donohue, T.M., Jr., Ganesan, M., Osna, N.A., and Kharbanda, K.K. (2021). Beneficial Effects of Betaine: A Comprehensive Review. Biology (Basel) *10*. 10.3390/biology10060456.
75. Seong, H.J., Roux, S., Hwang, C.Y., and Sul, W.J. (2022). Marine DNA methylation patterns are associated with microbial community composition and inform virus-host dynamics. Microbiome *10*, 157. 10.1186/s40168-022-01340-w.
76. Takahashi, M., Hiraoka, S., Matsumoto, Y., Shibagaki, R., Ujihara, T., Maeda, H., Seo, S., Nagasaki, K., Takeuchi, H., and Matsuzaki, S. (2025). Host-encoded DNA methyltransferases modify the epigenome and host tropism of invading phages. iScience *28*, 112264. 10.1016/j.isci.2025.112264.
77. Rani, A., Alam, A., Ahmad, F., P, M., Saurabh, A., Zarin, S., Mitra, D.K., Hasnain, S.E., and Ehtesham, N.Z. (2022). Mycobacterium tuberculosis Methyltransferase Rv1515c Can Suppress Host Defense Mechanisms by Modulating Immune Functions Utilizing a Multipronged Mechanism. Front Mol Biosci *9*, 906387. 10.3389/fmolb.2022.906387.
78. Kumar, V., and Tiwari, S.K. (2017). Halocin HA1: An archaeocin produced by the haloarchaeon Haloferax larsenii HA1. Process Biochemistry *61*, 202-208. https://doi.org/10.1016/j.procbio.2017.06.010.
79. Skennerton, C.T., Imelfort, M., and Tyson, G.W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res *41*, e105. 10.1093/nar/gkt183.
80. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics *8*, 209. 10.1186/1471-2105-8-209.
81. Steinegger, M., and Soding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol *35*, 1026-1028. 10.1038/nbt.3988.
82. Padilha, V.A., Alkhnbashi, O.S., Shah, S.A., de Carvalho, A., and Backofen, R.

(2020). CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. Gigascience *9*. 10.1093/gigascience/giaa062.

83. Camargo, A.P., Nayfach, S., Chen, I.A., Palaniappan, K., Ratner, A., Chu, K., Ritter, S.J., Reddy, T.B.K., Mukherjee, S., Schulz, F., et al. (2023). IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. Nucleic Acids Res *51*, D733-D743. 10.1093/nar/gkac1037.

84. Ahlgren, N.A., Chen, Y., Needham, D.M., Parada, A.E., Sachdeva, R., Trinh, V., Chen, T., and Fuhrman, J.A. (2017). Genome and epigenome of a novel marine Thaumarchaeota strain suggest viral infection, phosphorothioation DNA modification and multiple restriction systems. Environ Microbiol *19*, 2434-2452. 10.1111/1462-2920.13768.

85. Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. 10.1186/1471-2105-11-119.

86. Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Neron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res *46*, W246-W251. 10.1093/nar/gky425.

87. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol *39*, 578-585. 10.1038/s41587-020-00774-7.

88. Ren, J., Song, K., Deng, C., Ahlgren, N.A., Fuhrman, J.A., Li, Y., Xie, X., Poplin, R., and Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. Quant Biol *8*, 64-77. 10.1007/s40484-019-0187-4.

89. Camargo, A.P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P.S.G., Nayfach, S., and Kyrpides, N.C. (2023). Identification of mobile genetic elements with geNomad. Nat Biotechnol. 10.1038/s41587-023-01953-y.

90. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitua, M.C., Vik, D., Sullivan, M.B., and Roux, S. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 37. 10.1186/s40168-020-00990-y.

91. Hegarty, B., Riddell, V.J., Bastien, E., Langenfeld, K., Lindback, M., Saini, J.S., Wing, A., Zhang, J., and Duhaime, M. (2024). Benchmarking informatics

approaches for virus discovery: caution is needed when combining in silico identification methods. mSystems *9*, e0110523. 10.1128/msystems.01105-23.

92. Li, R., Wang, Y., Hu, H., Tan, Y., and Ma, Y. (2022). Metagenomic analysis reveals unexplored diversity of archaeal virome in the human gut. Nat Commun *13*, 7978. 10.1038/s41467-022-35735-y.

93. Chibani, C.M., Shah, S.A., Schmitz, R.A., and Nayfach, S. (2024). Inaccurate viral prediction leads to overestimated diversity of the archaeal virome in the human gut. Nat Commun *15*, 5976. 10.1038/s41467-024-49902-w.

94. Wang, Y., Li, R., and Ma, Y. (2024). Reply to: Inaccurate viral prediction leads to overestimated diversity of the archaeal virome in the human gut. Nat Commun *15*, 5977. 10.1038/s41467-024-49903-9.

95. Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425-430. 10.1038/nature19094.

96. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol *37*, 1530-1534. 10.1093/molbev/msaa015.

97. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504. 10.1101/gr.1239303.

98. Letunic, I., and Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res *52*, W78-W82. 10.1093/nar/gkae268.

99. Zheng, K., Sun, J., Liang, Y., Kong, L., Paez-Espino, D., McMinn, A., and Wang, M. (2025). VITAP: a high precision tool for DNA and RNA viral classification based on meta-omic data. Nat Commun *16*, 2226. 10.1038/s41467-025-57500-7.

100. Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res *40*, D130-135. 10.1093/nar/gkr1079.

101. von Meijenfeldt, F.A.B., Arkhipova, K., Cambuy, D.D., Coutinho, F.H., and Dutilh, B.E. (2019). Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome Biol *20*, 217. 10.1186/s13059-019-1817-x.

102. Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019).

Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol *37*, 632-639. 10.1038/s41587-019-0100-8.

103. Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics *36*, 2251-2252. 10.1093/bioinformatics/btz859.

104. Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladin, L., Raj, S., Richardson, L.J., et al. (2021). Pfam: The protein families database in 2021. Nucleic Acids Res *49*, D412-D419. 10.1093/nar/gkaa913.

105. Galperin, M.Y., Wolf, Y.I., Makarova, K.S., Vera Alvarez, R., Landsman, D., and Koonin, E.V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res *49*, D274-D281. 10.1093/nar/gkaa1018.

106. Darzi, Y., Letunic, I., Bork, P., and Yamada, T. (2018). iPath3.0: interactive pathways explorer v3. Nucleic Acids Res *46*, W510-W513. 10.1093/nar/gky299.

107. Kim, W., Mirdita, M., Levy Karin, E., Gilchrist, C.L.M., Schweke, H., Soding, J., Levy, E.D., and Steinegger, M. (2025). Rapid and sensitive protein complex alignment with Foldseek-Multimer. Nat Methods. 10.1038/s41592-025-02593-7.

108. Grazziotin, A.L., Koonin, E.V., and Kristensen, D.M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res *45*, D491-D498. 10.1093/nar/gkw975.

109. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol *61*, 539-542. 10.1093/sysbio/sys029.

110. Robinson, O., Dylus, D., and Dessimoz, C. (2016). Phylo.io: Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web. Mol Biol Evol *33*, 2163-2166. 10.1093/molbev/msw080.

111. Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res *41*, e121. 10.1093/nar/gkt263.

112. Mirdita, M., Schutze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat Methods *19*, 679-682. 10.1038/s41592-022-01488-1.

1065

Acknowledgements

We thank Jilu Han for his help with the installation and debugging of the software used in this study. We thank the support of the high-performance servers of the Center for High Performance Computing and System Simulation, Laoshan Laboratory (Qingdao), the High-Performance Biological Supercomputing Center at the Ocean University of China, the Marine Big Data Center of the Institute for Advanced Ocean Study of the Ocean University of China, and the IEMB-1, a high-performance computing cluster operated by the Institute of Evolution and Marine Biodiversity. This study was supported by the National Natural Science Foundation of China (No. 42120104006, 42176111, and 42306111), the Ocean Negative Carbon Emissions (ONCE), 2024 Graduate Self-directed Research Project (2024ZZKY, 202461036), and the Fundamental Research Funds for the Central Universities (202172002, 201812002, 202461036, Andrew McMinn).

Author contributions

Conceptualization: K.Z. and Y.L.; Analysis: K.Z.; Database: D.P.-E.; Writing manuscript: K.Z.; Review and editing: Y.L., A.M., and M.W.; Computational resource: D.P.-E. and M.W.; Funding acquisition and supervision: Y.L., A.M., and M.W.

Competing interests

The authors declare no competing interests for the manuscript “Viruses associated with *Nitrososphaerota* archaea: diversity, evolution and horizontal gene transfers”.

Figure legends

Figure 1 The establishment of *Nitrososphaerota*-associated viruses and mobile genetic elements (MGEs). a, The workflow for mining, filtering, and identifying viruses/MGEs primarily included the prediction of archaeal CRISPR systems, viral/MGE mining, genome filtering, and the definition of viral operational taxonomic units (vOTUs). Only genomes encoding detectable capsid or related proteins were considered viruses; others were classified as MGEs. b, The data sources for *Nitrososphaerota*-associated viruses/MGEs included in this genomic catalogue. Due to the pipeline's prioritization of specificity, not all previously reported *Nitrososphaerota*-associated viral/MGE genomes were included. c, Habitat sources of Nitrososphaerota-associated viruses/MGEs in this genomic catalogue at the vOTU level.

Figure 2 Overview of the diversity of the genome catalogue of *Nitrososphaerota*-associated viruses and mobile genetic elements (MGEs). a, Spacer-sharing network among *Nitrososphaerota* genomes. Nodes represent archaeal genomes and are colored by distinct archaeal lineages. An edge is generated between two genomes if they share spacer sequences, suggesting that their CRISPR systems may target the same viral/MGE species. Bold edges indicate that the shared spacer sequences have corresponding protospacers present in the *Nitrososphaerota*-associated viral/MGE catalogue generated in this study. b, Biogeographic distribution of *Nitrososphaerota*-associated viral/MGE genomes. In the map, nodes of different colors indicate their putative archaeal hosts: *Aigarchaea*, ammonia-oxidizing archaea (AOA), non-ammonia-oxidizing *Thaumarchaea* (Non-AO Thaum), and deeply-branching *Nitrososphaerota* archaea (D-B NitroA). c, The number of viral operational taxonomic units (vOTUs) from each habitat in the *Nitrososphaerota*-associated viral/MGE catalogue. Bar colors represent different archaeal lineages, including Group 1.1a,

Nitrosotales, Group 1.1b, Marine Trench and Ground Water (MT/GW), *Nitrosocaldales*, Marine Beneath Group A (MBG-A), deeply-branching *Thaumarchaea* (D-B Thaum), *Wolframiiraptoraceae*, *Caldarchaeales*, pSL4-affiliation, and D-B NitroA. Virion icons to the right of the bar plot illustrate the numbers of viruses/MGEs with different possible morphologies across habitats. d, Phylogenetic tree of *Nitrososphaerota* archaeal genomes. This tree was constructed based on concatenated alignments of 46 ribosomal proteins (totaling 5,963 amino acid positions) using the maximum likelihood method with the LG+C60+F model as recommended by Spang et al.[2] (see Methods). Reference genomes from *Thermoprotei* were used as the outgroup to root the tree. Tree annotation includes the habitats and lineages of these archaeal genomes. Virion icons to the right of the tree indicate the numbers of viruses/MGEs with different possible morphologies (at the vOTU level) associated with each archaeal lineage.

Figure 3 Similarity of *Nitrososphaerota*-associated viruses and mobile genetic elements (MGEs) based on their genome contents. a, Genome-content similarity network of *Nitrososphaerota*-associated viruses/MGEs. Nodes represent viral/MGE genomes and are coloured according to their associated archaeal lineages. Each distinct super viral cluster (sVC) may consist of one or more viral clusters (VCs). The percentage of shared proteins between genomes is indicated by the thickness of the edges connecting the nodes. The shapes of the nodes denote the potential morphologies of these viruses/MGEs. b, Viral proteomic tree used for the delineation of viral families. A distance threshold of 0.05 and a minimum of six shared proteins among members were used as the criteria for defining families [27] (see Methods). The names of the proposed viral families or MGE families are indicated on the right side of the tree, with different virion icons representing the possible morphologies of these families. Bar plots indicate the percentage of shared proteins within each family. Tree annotations, from left to right, include genome length, G+C content (%), coding density (%), habitats, host lineage, and morphology. Abbreviations: MT/GW, Marine Trench and Ground Water; MBG-A, Marine Beneath Group A; D-B Thaum, deeply-

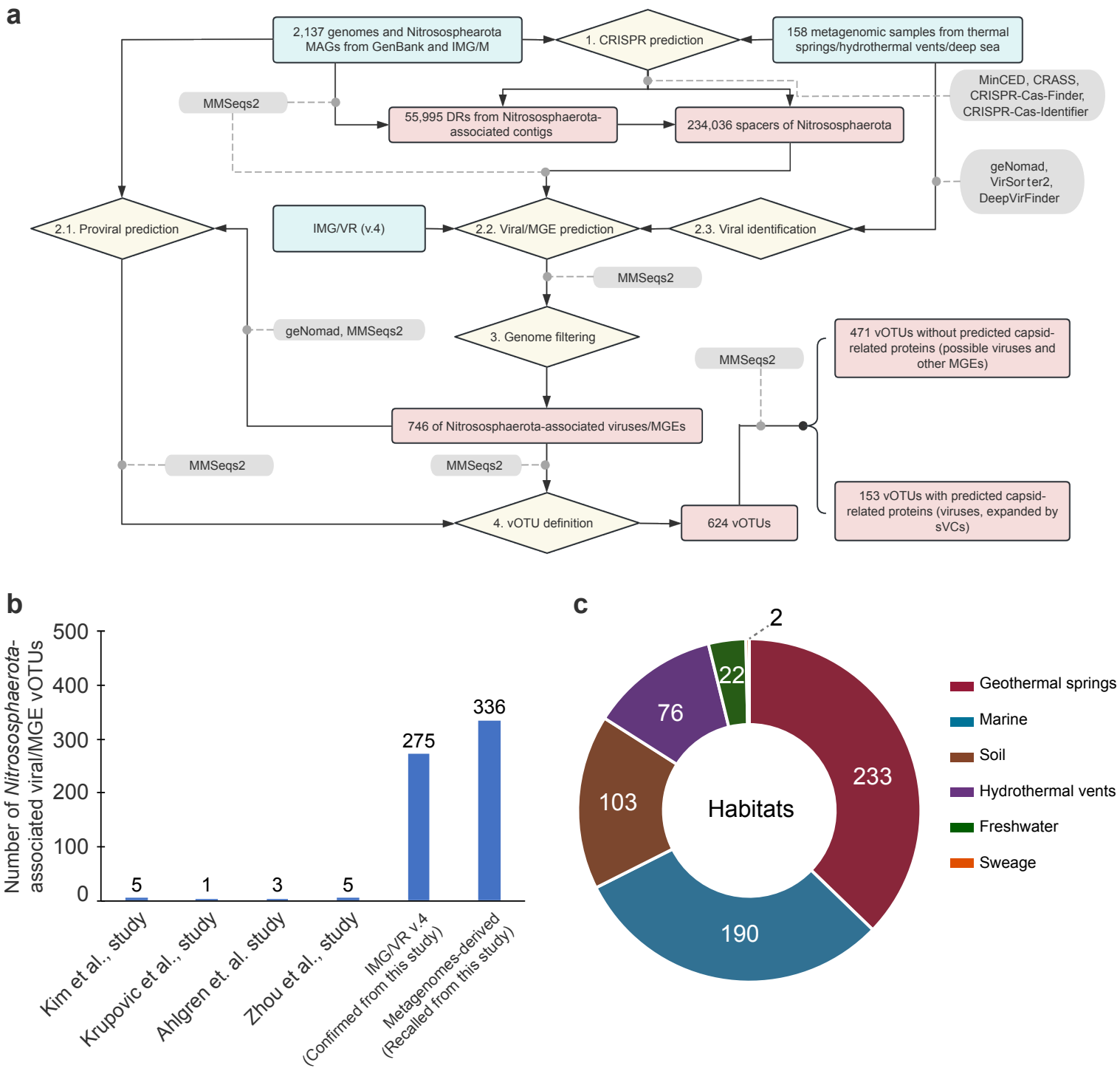
branching Thaumarchaea; D-B NitroA, deeply-branching Nitrososphaerota archaea.

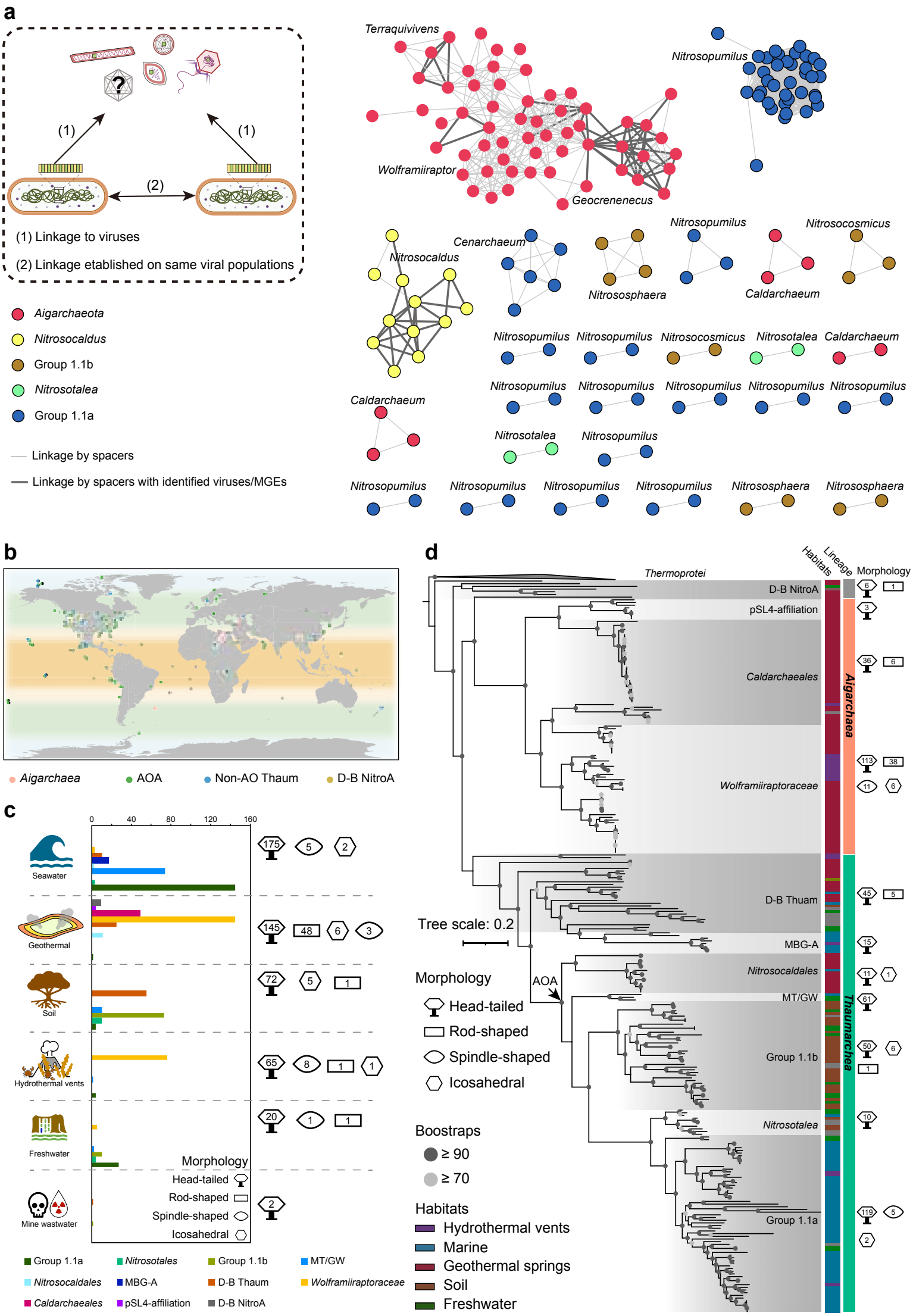
Figure 4 Phylogenetic inferences of *Nitrososphaerota*-associated viruses and mobile genetic elements (MGEs) based on viral hallmark proteins (VHPs) through Bayesian inference (BI) and Maximum Likelihood (ML) joint approach. a, The trees display the topologies inferred by BI based on VHPs. *Nitrososphaerota*-associated viruses/MGEs are highlighted with bold red branches. To the right of each tree are shown: 1) BI convergence indices, including minimum effective sample size (minESS), potential scale reduction factor (PSRF), and average standard deviation of split frequencies (ASDSF); 2) indices for comparing BI and ML topologies, including normalized Robinson–Foulds (RF) distance based on ML and the number of clades shared between BI and ML. Posterior probabilities from BI are indicated on the tree branches with colored nodes, alongside ML bootstrap values (MLB). Names of viral families (in bold red font) and other host groups are annotated beside the corresponding branches. Tree annotations indicate the *Nitrososphaerota* lineages associated with these viruses/MGEs. a, Bayesian tree based on the SIRV2-like capsid protein (domain accession in Pfam: PF12193). DUF732 (domain accession in Pfam: PF05305) was used as the outgroup to root the tree. Leaves corresponding to rod-shaped viruses with heterodimeric capsids are highlighted with red squares. b, Bayesian tree based on the HK97-like capsid protein (domain accession in Pfam: PF05065), rooted using L_lactis_ph-MCP (domain accession in Pfam: PF06673) as the outgroup. c, Bayesian tree based on the HK97-like capsid protein (domain accession in VOGdb: VOG01305), rooted using C1_1 (domain accession in Pfam: PF00130) as the outgroup. d, Bayesian tree based on the HK97-like capsid protein (domain accession in Pfam: PF07068), rooted using L_lactis_ph-MCP (domain accession in Pfam: PF06673) as the outgroup. e, Bayesian tree based on the terminase large subunit protein (domain accession in KEGG: K06909), rooted using P-loop NTPase (domain accession in SUPERFAMILY: SSF52540) as the outgroup. f, Bayesian tree based on the terminase large subunit protein

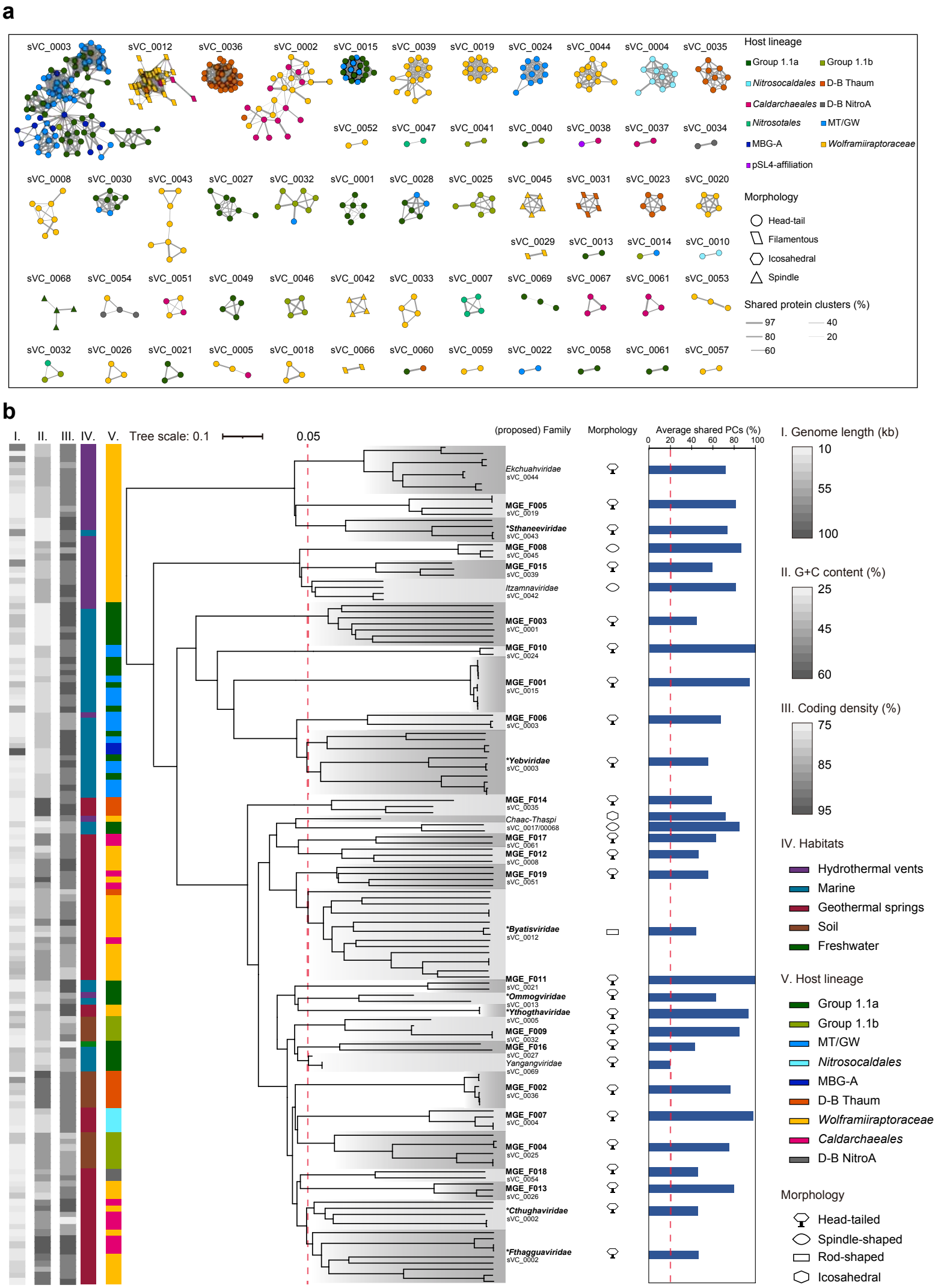
(domain accession in Pfam: PF03237), rooted using P-loop NTPase (domain accession in SUPERFAMILY: SSF52540) as the outgroup. Abbreviations: MT/GW, Marine Trench and Ground Water; MBG-A, Marine Beneath Group A; D-B Thaum, deeply-branching *Thaumarchaea*; D-B NitroA, deeply-branching *Nitrososphaerota* archaea.

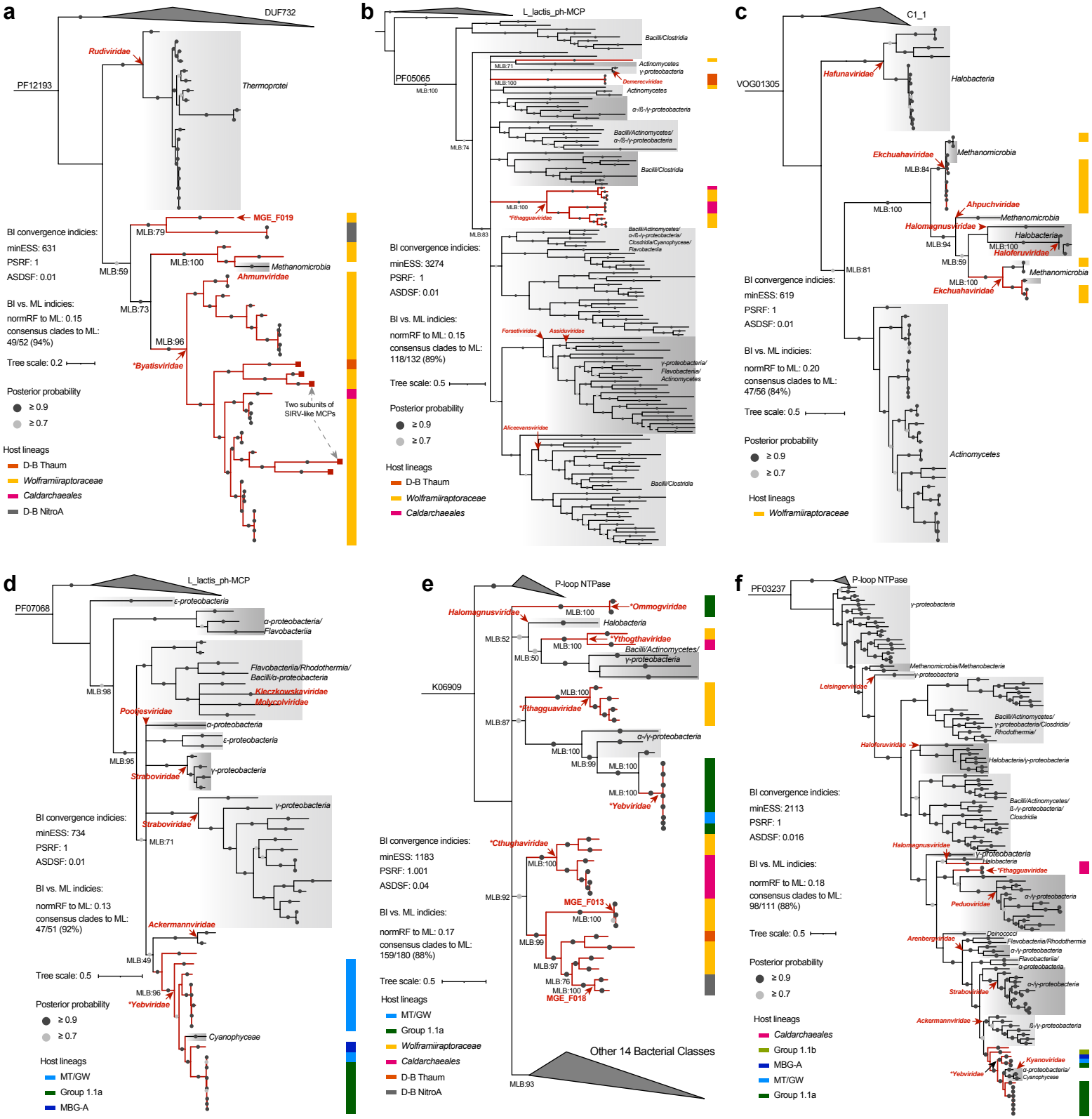
Figure 5 Genes encoded by *Nitrososphaerota*-associated viruses and mobile genetic elements (MGEs) may be involved in metabolic compensation for their archaeal hosts. Hexagons of different colours represent viruses/MGEs associated with different *Nitrososphaerota* lineages. If a particular metabolic gene is found only in viral/MGE genomes but is absent from the panproteome of the corresponding archaeal lineage, it is considered as compensation from viruses/MGEs and is highlighted with an asterisk. These metabolic genes were assigned to 11 different metabolic categories based on the KEGG database (Supplementary Dataset 7). Different metabolic categories based on KEGG are indicated in bold next to the corresponding modules.

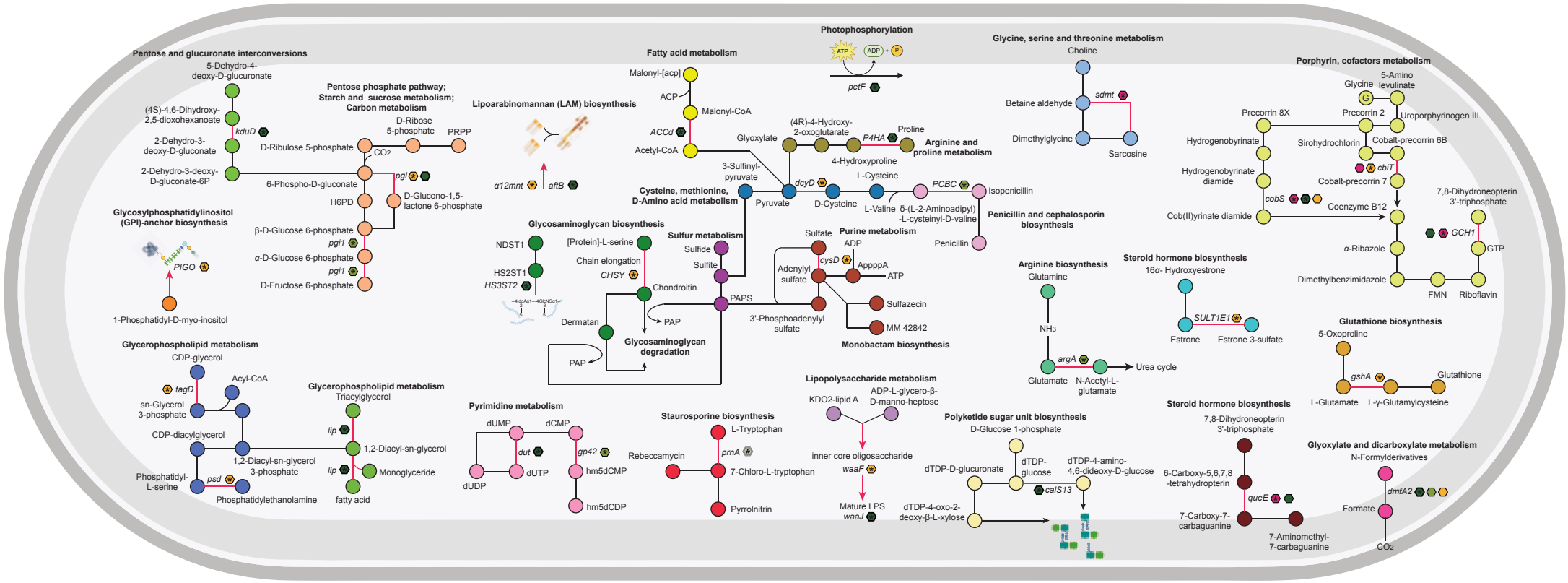
Figure 6 Metabolic potential and horizontal gene transfer (HGT) landscape of *Nitrososphaerota*-associated viruses and mobile genetic elements (MGEs). a, Metabolic genes encoded by viral/MGE genomes but not detected in their corresponding hosts. These metabolic genes were assigned to 11 different metabolic categories based on the KEGG database (Supplementary Dataset 7). b, Proportion of HGT events with different transfer directions among various archaeal lineages. c, Relationships among HGT event direction, archaeal lineages, and viral/MGE families. The number of events is indicated in parentheses. d, Relationships among HGT event direction, viral function, cellular function, and elements-related metabolism.


a
2,137 genomes and Nitrososphearota MAGs from GenBank and IMG/M
1. CRISPR prediction
158 metagenomic samples from thermal springs/hydrothermal vents/deep sea
MMSeqs2
MinCED, CRASS, CRISPR-Cas-Finder, CRISPR-Cas-Identifier
55,995 DRs from Nitrososphaerota-associated contigs
234,036 spacers of Nitrososphaerota
geNomad, VirSorter2, DeepVirFinder
2.1. Proviral prediction
IMG/VR (v.4)
2.2. Viral/MGE prediction
2.3. Viral identification
MMSeqs2
geNomad, MMSeqs2
3. Genome filtering
MMSeqs2
471 vOTUs without predicted capsid-related proteins (possible viruses and other MGEs)
746 of Nitrososphaerota-associated viruses/MGEs
MMSeqs2
MMSeqs2
153 vOTUs with predicted capsid-related proteins (viruses, expanded by sVCs)
4. vOTU definition
624 vOTUs
b
Number of Nitrososphaerota-associated viral/MGE vOTUs
500
400
300
200
100
0
5
1
3
5
275
336
Kim et al., study
Krupovic et al., study
Ahlgren et. al. study
Zhou et al., study
IMG/VR v.4 (Confirmed from this study)
Metagenomes-derived (Recalled from this study)
c
2
22
76
103
233
190
Habitats
Geothermal springs
Marine
Soil
Hydrothermal vents
Freshwater
Sweage


a
(1)
(1)
(2)
(1) Linkage to viruses
(2) Linkage etablished on same viral populations
Aigarchaeota
Nitrosocaldus
Group 1.1b
Nitrosotalea
Group 1.1a
Linkage by spacers
Linkage by spacers with identified viruses/MGEs
Terraquivivens
Nitrosopumilus
Wolframiiraptor
Geocrenenecus
Nitrosocaldus
Cenarchaeum
Nitrosopumilus
Nitrosocosmicus
Nitrososphaera
Caldarchaeum
Nitrosopumilus
Nitrosopumilus
Nitrosocosmicus
Nitrosotalea
Caldarchaeum
Nitrosopumilus
Nitrosopumilus
Nitrosopumilus
Nitrosopumilus
Nitrosopumilus
Caldarchaeum
Nitrosotalea
Nitrosopumilus
Nitrosopumilus
Nitrosopumilus
Nitrosopumilus
Nitrosopumilus
Nitrososphaera
Nitrososphaera
b
Aigarchaea
AOA
Non-AO Thaum
D-B NitroA
c
0
40
80
120
160
Seawater
175
5
2
Geothermal
145
48
6
3
Soil
72
5
1
Hydrothermal vents
65
8
1
1
Freshwater
20
1
1
Mine wastwater
2
Morphology
Head-tailed
Rod-shaped
Spindle-shaped
Icosahedral
Group 1.1a
Nitrosotales
Group 1.1b
MT/GW
Nitrosocaldales
MBG-A
D-B Thaum
Wolframiiraptoraceae
Caldarchaeales
pSL4-affiliation
D-B NitroA
d
Thermoprotei
Habitats
Lineage
Morphology
D-B NitroA
pSL4-affiliation
Caldarchaeales
Wolframiiraptoraceae
D-B Thuam
MBG-A
Nitrosocaldales
MT/GW
Group 1.1b
Nitrosotalea
Group 1.1a
Aigarchaea
Thaumarchea
Tree scale: 0.2
AOA
Morphology
Head-tailed
Rod-shaped
Spindle-shaped
Icosahedral
Boostraps
≥ 90
≥ 70
Habitats
Hydrothermal vents
Marine
Geothermal springs
Soil
Freshwater


a
sVC_0003
sVC_0012
sVC_0036
sVC_0002
sVC_0015
sVC_0039
sVC_0019
sVC_0024
sVC_0044
sVC_0004
sVC_0035
sVC_0052
sVC_0047
sVC_0041
sVC_0040
sVC_0038
sVC_0037
sVC_0034
sVC_0008
sVC_0030
sVC_0043
sVC_0027
sVC_0032
sVC_0001
sVC_0028
sVC_0025
sVC_0045
sVC_0031
sVC_0023
sVC_0020
sVC_0029
sVC_0013
sVC_0014
sVC_0010
sVC_0068
sVC_0054
sVC_0051
sVC_0049
sVC_0046
sVC_0042
sVC_0033
sVC_0007
sVC_0069
sVC_0067
sVC_0061
sVC_0053
sVC_0032
sVC_0026
sVC_0021
sVC_0005
sVC_0018
sVC_0066
sVC_0060
sVC_0059
sVC_0022
sVC_0058
sVC_0061
sVC_0057
Host lineage
Group 1.1a
Group 1.1b
Nitrosocaldales
D-B Thaum
Caldarchaeales
D-B NitroA
Nitrosotales
MT/GW
MBG-A
Wolframiiraptoraceae
pSL4-affiliation
Morphology
Head-tail
Filamentous
Icosahedral
Spindle
Shared protein clusters (%)
97
80
60
40
20
b
I.
II.
III.
IV.
V.
Tree scale: 0.1
0.05
(proposed) Family
Morphology
Average shared PCs (%)
0
20
40
60
80
100
Ekchuahviridae
sVC_0044
MGE_F005
sVC_0019
*Sthaneeviridae
sVC_0043
MGE_F008
sVC_0045
MGE_F015
sVC_0039
Itzamnaviridae
sVC_0042
MGE_F003
sVC_0001
MGE_F010
sVC_0024
MGE_F001
sVC_0015
MGE_F006
sVC_0003
*Yebviridae
sVC_0003
MGE_F014
sVC_0035
Chaac-Thaspi
sVC_0017/00068
MGE_F017
sVC_0061
MGE_F012
sVC_0008
MGE_F019
sVC_0051
*Byatisviridae
sVC_0012
MGE_F011
sVC_0021
*Ommogviridae
sVC_0013
*Ythogthaviridae
sVC_0005
MGE_F009
sVC_0032
MGE_F016
sVC_0027
Yangangviridae
sVC_0069
MGE_F002
sVC_0036
MGE_F007
sVC_0004
MGE_F004
sVC_0025
MGE_F018
sVC_0054
MGE_F013
sVC_0026
*Cthughaviridae
sVC_0002
*Fthagguaviridae
sVC_0002
I. Genome length (kb)
10
55
100
II. G+C content (%)
25
45
60
III. Coding density (%)
75
85
95
IV. Habitats
Hydrothermal vents
Marine
Geothermal springs
Soil
Freshwater
V. Host lineage
Group 1.1a
Group 1.1b
MT/GW
Nitrosocaldales
MBG-A
D-B Thaum
Wolframiiraptoraceae
Caldarchaeales
D-B NitroA
Morphology
Head-tailed
Spindle-shaped
Rod-shaped
Icosahedral


a
DUF732
PF12193
Rudiviridae
Thermoprotei
MGE_F019
MLB:79
MLB:59
MLB:100
Methanomicrobia
Ahmunviridae
MLB:73
MLB:96
*Byatisviridae
Two subunits of SIRV-like MCPs
BI convergence indicies:
minESS: 631
PSRF: 1
ASDSF: 0.01
BI vs. ML indicies:
normRF to ML: 0.15
consensus clades to ML: 49/52 (94%)
Tree scale: 0.2
Posterior probability
≥ 0.9
≥ 0.7
Host lineages
D-B Thaum
Wolframiiraptoraceae
Caldarchaeales
D-B NitroA
b
L_lactis_ph-MCP
PF05065
MLB:100
Bacilli/Clostridia
MLB:71
Actinomycetes
γ-proteobacteria
Oemerecoviridae
MLB:100
Actinomycetes
α-/ß-/γ-proteobacteria
Bacilli/Actinomycetes/
α-/ß-/γ-proteobacteria
MLB:74
Bacilli/Clostridia
MLB:100
*Fthagguaviridae
MLB:83
Bacilli/Actinomycetes/
α-/ß-/γ-proteobacteria/
Clostridia/Cyanophyceae/
Flavobacteria
Forsetiviridae
Assiduviridae
γ-proteobacteria/
Flavobacteria/
Actinomycetes
Aliceevansviridae
Bacilli/Clostridia
BI convergence indicies:
minESS: 3274
PSRF: 1
ASDSF: 0.01
BI vs. ML indicies:
normRF to ML: 0.15
consensus clades to ML: 118/132 (89%)
Tree scale: 0.5
Posterior probability
≥ 0.9
≥ 0.7
Host lineages
D-B Thaum
Wolframiiraptoraceae
Caldarchaeales
c
C1_1
VOG01305
Hafunaviridae
Halobacteria
Ekchuahaviridae
Methanomicrobia
MLB:84
MLB:100
Ahpuchviridae
Methanomicrobia
Halomagnusviridae
Halobacteria
MLB:94
Haloferuviridae
MLB:59
MLB:81
Methanomicrobia
MLB:100
Ekchuahaviridae
Actinomycetes
BI convergence indicies:
minESS: 619
PSRF: 1
ASDSF: 0.01
BI vs. ML indicies:
normRF to ML: 0.20
consensus clades to ML: 47/56 (84%)
Tree scale: 0.5
Posterior probability
≥ 0.9
≥ 0.7
Host lineages
Wolframiiraptoraceae
d
L_lactis_ph-MCP
PF07068
ε-proteobacteria
α-proteobacteria/
Flavobacteria
MLB:98
Flavobacteria/Rhodothermia/
Bacilli/α-proteobacteria
Klaczkowskaviridae
Malyovolviridae
Poolesviridae
α-proteobacteria
ε-proteobacteria
MLB:95
Straboviridae
γ-proteobacteria
γ-proteobacteria
Straboviridae
MLB:71
Ackermannviridae
MLB:49
MLB:96
*Yebviridae
Cyanophyceae
BI convergence indicies:
minESS: 734
PSRF: 1
ASDSF: 0.01
BI vs. ML indicies:
normRF to ML: 0.13
consensus clades to ML: 47/51 (92%)
Tree scale: 0.5
Posterior probability
≥ 0.9
≥ 0.7
Host lineages
MT/GW
Group 1.1a
MBG-A
e
P-loop NTPase
K06909
Halomagnusviridae
MLB:100
*Ommogviridae
Halobacteria
MLB:52
MLB:100
*Ythogthaviridae
MLB:50
Bacilli/Actinomycetes/
γ-proteobacteria
MLB:100
*Fthagguaviridae
MLB:87
α-/γ-proteobacteria
MLB:100
MLB:99
MLB:100
MLB:100
*Yebviridae
*Cthughaviridae
MLB:100
MLB:92
MGE_F013
MLB:100
MLB:99
MLB:97
MLB:76
MLB:100
MGE_F018
Other 14 Bacterial Classes
MLB:93
BI convergence indicies:
minESS: 1183
PSRF: 1.001
ASDSF: 0.04
BI vs. ML indicies:
normRF to ML: 0.17
consensus clades to ML: 159/180 (88%)
Tree scale: 0.5
Host lineages
MT/GW
Group 1.1a
Wolframiiraptoraceae
Caldarchaeales
D-B Thaum
D-B NitroA
f
PF03237
P-loop NTPase
γ-proteobacteria
Methanomicrobia/Methanobacteria
γ-proteobacteria
Leisingerviridae
Bacilli/Actinomycetes/
γ-proteobacteria/Clostridia/
Rhodothermia/
Haloferuviridae
Halobacteria/γ-proteobacteria
Bacilli/Actinomycetes/
ß-γ-proteobacteria/
Clostridia
Halomagnusviridae
γ-proteobacteria
Halobacteria
*Fthagguaviridae
α-γ-proteobacteria
Peduoviridae
Deinococci
Arenbergviridae
Flavobacteria/Rhodothermia
α-γ-proteobacteria
Flavobacteria/
α-proteobacteria
α-γ-proteobacteria
Straboviridae
Ackermannviridae
ß-γ-proteobacteria
Kyanoviridae
α-proteobacteria/
Cyanophyceae
*Yebviridae
BI convergence indicies:
minESS: 2113
PSRF: 1
ASDSF: 0.016
BI vs. ML indicies:
normRF to ML: 0.18
consensus clades to ML: 98/111 (88%)
Tree scale: 0.5
Posterior probability
≥ 0.9
≥ 0.7
Host lineages
Caldarchaeales
Group 1.1b
MBG-A
MT/GW
Group 1.1a



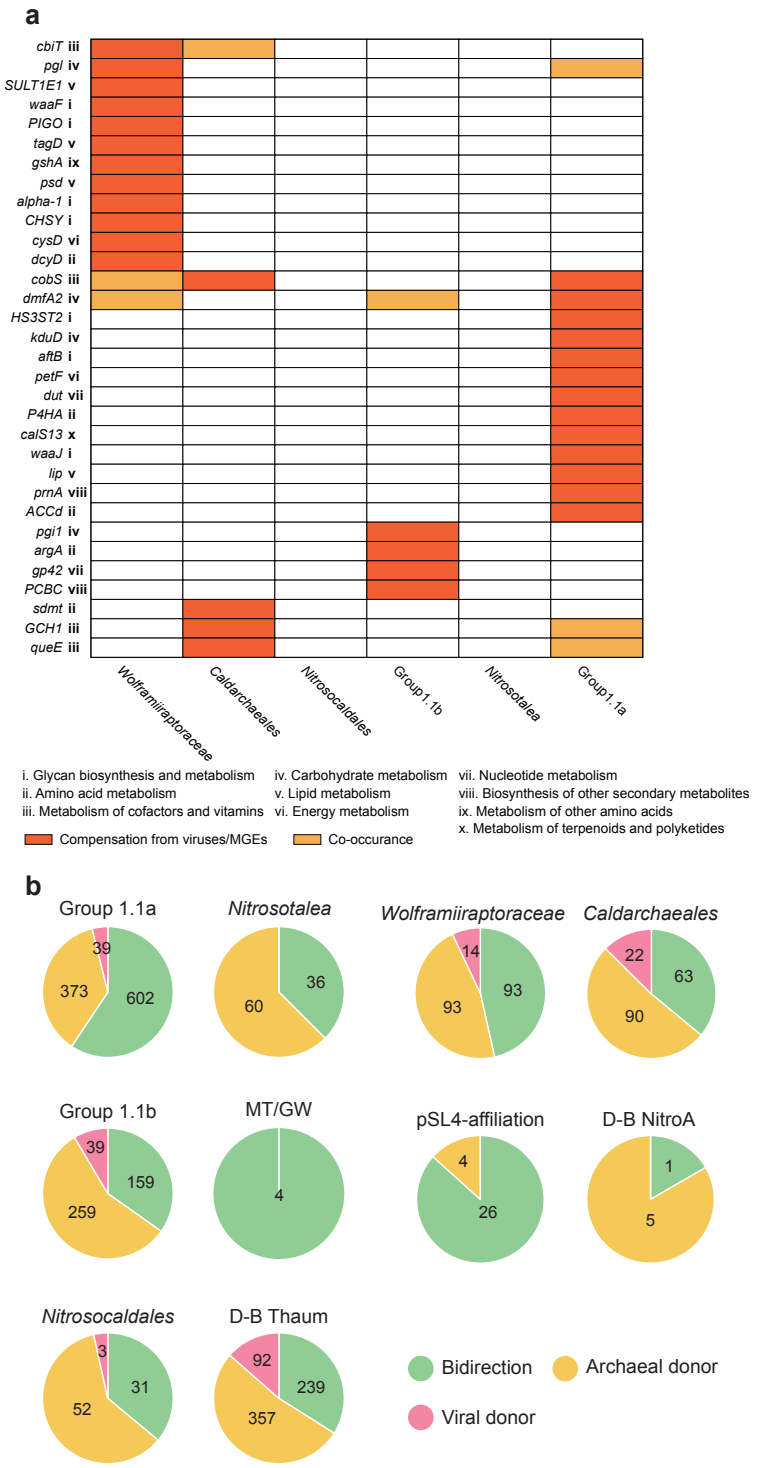
Wolframiiraptoraceae

Caldarchaeales

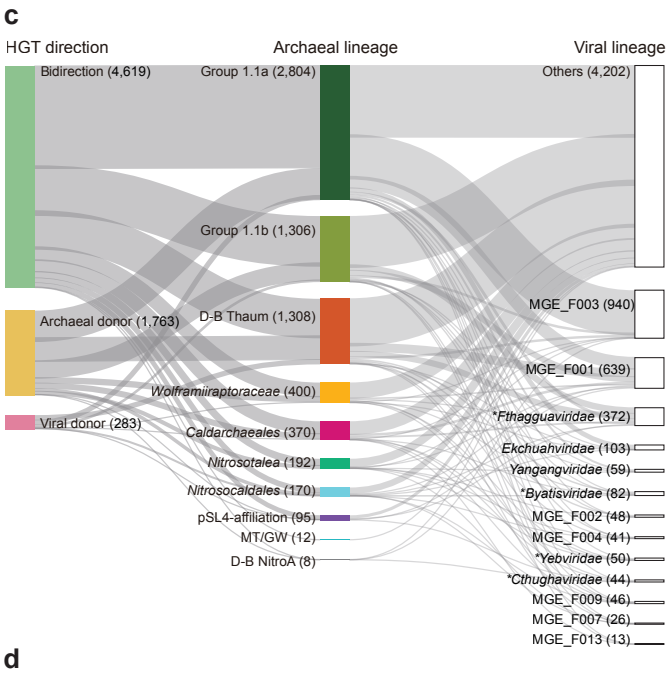
Group 1.1b

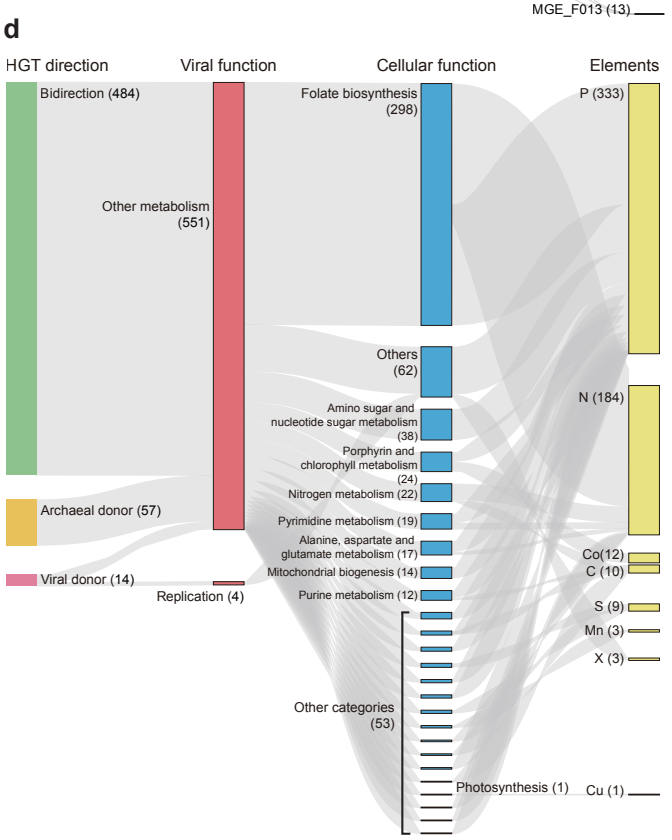
Group 1.1a

* Compensation from viruses/MGEs


a
cbiT iii
pgl iv
SULT1E1 v
waaF i
PIGO i
tagD v
gshA ix
psd v
alpha-1 i
CHSY i
cysD vi
dcyD ii
cobS iii
dmfA2 iv
HS3ST2 i
kduD iv
aftB i
petF vi
dut vii
P4HA ii
calS13 x
waaJ i
lip v
prnA viii
ACCd ii
pgi1 iv
argA ii
gp42 vii
PCBC viii
sdmt ii
GCH1 iii
queE iii
Wolframiiraptoraceae
Caldarchaeales
Nitrosocaldales
Group1.1b
Nitrosotalea
Group1.1a
i. Glycan biosynthesis and metabolism
ii. Amino acid metabolism
iii. Metabolism of cofactors and vitamins
iv. Carbohydrate metabolism
v. Lipid metabolism
vi. Energy metabolism
vii. Nucleotide metabolism
viii. Biosynthesis of other secondary metabolites
ix. Metabolism of other amino acids
x. Metabolism of terpenoids and polyketides
Compensation from viruses/MGEs
Co-occurance

b
Group 1.1a
39
373
602
Nitrosotalea
36
60
Wolframiiraptoraceae
14
93
93
Caldarchaeales
22
63
90
Group 1.1b
39
159
259
MT/GW
4
pSL4-affiliation
4
26
D-B NitroA
1
5
Nitrosocaldales
3
31
52
D-B Thaum
92
239
357
Bidirection
Archaeal donor
Viral donor


c
HGT direction
Archaeal lineage
Viral lineage
Bidirection (4,619)
Archaeal donor (1,763)
Viral donor (283)
Group 1.1a (2,804)
Group 1.1b (1,306)
D-B Thaum (1,308)
Wolframiiraptoraceae (400)
Caldarchaeales (370)
Nitrosotalea (192)
Nitrosocaldales (170)
pSL4-affiliation (95)
MT/GW (12)
D-B NitroA (8)
Others (4,202)
MGE_F003 (940)
MGE_F001 (639)
*Fthagguaviridae (372)
Ekchuahviridae (103)
Yangangviridae (59)
*Byatisviridae (82)
MGE_F002 (48)
MGE_F004 (41)
*Yebviridae (50)
*Cthughaviridae (44)
MGE_F009 (46)
MGE_F007 (26)
MGE_F013 (13)


d
HGT direction
Viral function
Cellular function
Elements
Bidirection (484)
Archaeal donor (57)
Viral donor (14)
Other metabolism (551)
Replication (4)
Folate biosynthesis (298)
Others (62)
Amino sugar and nucleotide sugar metabolism (38)
Porphyrin and chlorophyll metabolism (24)
Nitrogen metabolism (22)
Pyrimidine metabolism (19)
Alanine, aspartate and glutamate metabolism (17)
Mitochondrial biogenesis (14)
Purine metabolism (12)
Other categories (53)
Photosynthesis (1)
P (333)
N (184)
Co(12)
C (10)
S (9)
Mn (3)
X (3)
Cu (1)