Expanding diversity of virome associated with ammonium-oxidizing archaea and related lineages

Kaiyang Zheng[1], Yantao Liang[1,2,3*], Fabai Wu[4*], Jiaxiang Li[1], David Paez-Espino[5], Andrew McMinn[1,3,6*], Min Wang[1,2,3,7*]

[1] College of Marine Life Sciences, MoE Key Laboratory of Evolution & Marine Biodiversity, Institute of Evolution and Marine Biodiversity, Frontiers Science Centre for Deep Ocean Multispheres and Earth System, Centre for Ocean Carbon Neutrality, Ocean University of China, Qingdao, China.

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

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

[4] Faculty of Science, Eastern Institute of Technology, Ningbo, China

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

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

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

* Correspondence:

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

Fabai Wu, Email: **fabaiwu@eitech.edu.cn**

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

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

Abstract

Nitrososphaeria, an archaeal class containing all ammonium-oxidizing archaea (AOA) and their closely related clades (e.g. *Ca. Caldarchaeales*), is a widely distributed archaeal lineage across diverse ecosystems. Due to the difficulty of isolation and cultivation, the viral diversity and evolution associated with non-marine AOA lineages and *Ca. Caldarchaeales* remain largely uncharacterized. Here, based on globally sampled genomes and metagenomes, we identified 57 and 230 species-level viruses or mobile genetic elements (MGEs) associated with AOA, including 82 sequences longer than 10 kb. Importantly, 85 and 178 previously unrecognized species-level viruses and MGEs associated with *Ca. Caldarchaeales* were identified, including 85 sequences longer than 10 kb. AOA-associated viruses were classified into four viral families, with one representing newly identified head-tailed lineages; *Caldarchaeales*-associated viruses were classified into nine families, including four novel head-tailed and one novel rod-shaped family. Phylogenetic inference revealed that the *Caldarchaeales*-associated rod-shaped viral family (*Demeterviridae*) shares a common ancestor with rod-shaped viruses of methanogens and they exhibit a unique phenomenon of major capsid gene duplication. Functional genomics analyses further revealed that these novel rod-shaped viruses encode virus-specific rare quinone biosynthesis genes and arsenic response transcriptional regulator genes, providing potential clues to their mechanisms of adaptation to hyperthermal environments. Collectively, our work uncovers the hidden genomic, phylogenetic and functional diversity of *Nitrososphaeria* viromes for the first time and closes a long-standing gap in archaeal virome research.

Introduction

The *Nitrososphaeria* is an archaeal class within *Thermoprotei*, encompassing the previously defined *Thaumarchaeota* and *Augarchaeota*, both of which are important members of the *Thermoproteati* (previously known as TACK superphylum)[1,2]. In many aquatic ecosystems, *Nitrososphaeria* constitute up to 40% of microorganisms, contributing significantly to ammonia oxidation, thus heavily influencing nitrogen cycling [3-5]. *Nitrososphaeria* includes all ammonia-oxidizing archaea (AOA) [6] and utilizes 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]. The ancestors of AOA are considered to have originated in terrestrial environments and to have adapted aerobically after the Great Oxygenation Event [13,15,16]. Subsequently, these archaea became key microbial players across diverse ecosystems and make a substantial contribution to carbon fixation and ammonia oxidation [8,15,17-21]. As the sister clade of AOA, *Ca. Caldarchaeales* shares substantial genetic content involved in sulphur oxidation and ammonia oxidation, suggesting an ancestral ecological transition of AOA from hyperthermal to moderate environments [22]. Primarily inhabiting hydrothermal vents and hot springs, *Ca. Caldarchaeales* influences carbon and sulphur cycling, nutrient transformation and microbial community structures, playing critical ecological roles in extreme environments [23,24].

Viruses and mobile genetic elements (MGEs) exhibit complex morphological and genomic diversity, significantly shaping archaeal genomes and their ecological footprint [25,26]. Due to cultivation challenges, environmental metagenomics has become the primary method for exploring archaeal viral diversity. Recent studies using environmental metagenomes have greatly expanded viral diversity associated with the Asgard super-phylum, methanogens and DPANN super-phylum, addressing gaps in viral evolution and host-virus interactions [27-31]. Over 50 AOA-associated viral or MGE sequences from marine samples have so far been identified, revealing diverse

viral elements that influence ammonia oxidation and host genomic organization [32-38]. However, the limited number of case studies prevents a comprehensive profiling of the diversity, evolution and functions of these archaeal viruses. The absence of viruses associated with extremophilic AOA lineages (e.g., *Nitrosocaldales* and *Nitrosotalea*) poses a challenge for resolving phylogenetic placements of AOA viruses within the virosphere. In particular, as a phylogenetic sister clade closely related to AOA and one of the main lineages constituting the TACK super-phylum, *Caldarchaeales*-associated viruses or MEGs remain underexplored and represent a major enigma within virology.

To address this knowledge gap, we systematically explored *Nitrososphaeria*-associated viruses or MGEs diversity based on extensive global genomic and metagenomic datasets. Unlike previous studies that focused only on viruses from a single AOA lineage [32-36], a comprehensive analysis of the *Nitrososphaeria*-associated virome was conducted, yielding systematic insights into its genomics, phylogenetics and virus–host interactions. Through the establishment of a genomic catalogue comprising 746 viral or MGE sequences (including 624 species-level sequences), we revealed six previously unrecognized archaeal viral families, including a group of rod-shaped viruses associated with *Ca. Caldarchaeales*, thereby providing new insights into the archaeal virosphere.

Results

Overview of the genomic catalogue of *Nitrososphaeria*-associated viruses or MGEs

Nitrososphaeria genome database containing 2,137 genomes and metagenome-assembled genomes (MAGs) was constructed from NCBI GenBank and IMG/M databases (Fig. 1a and Supplementary Note 1)[39,40]. The virus or MGE database comprised over 15 million contigs from IMG/VR v.4, together with 170 metagenome-assembled contigs originating from hyperthermal, abyssal/hadal ocean, hypersaline, frozen, and acidic environments, including 17 from our published and unpublished studies (Supplementary Dataset 1) [39,41-44]. By integrating CRISPR spacer matches and protein-sharing analysis, 746 *Nitrososphaeria*-associated viruses or MGEs were identified, including 712 extrachromosomal and 34 integrated ones. These were grouped into 624 species-level viral operative taxonomic units (vOTUs), comprising 153 vOTUs with and 471 MGE OTUs without known viral capsid proteins (Fig. 1a, Supplementary Dataset 2). In addition, one vOTU and 39 MGE OTUs associated with AOA and 79 vOTUs and 175 MGE OTUs associated with *Ca. Caldarchaeales* were newly identified, respectively.

Nitrososphaeria-associated viruses or MGEs are widely distributed from tropical to polar regions and across terrestrial to marine environments (Fig. 1b). Most of them originated from geothermal, marine, soil and hydrothermal environments, aligning with habitats of AOA, deeply-branched AOA-related lineage and *Ca. Caldarchaeales* (Fig. 1b) [7,20,22,23,45-48]. Overall, this genome catalogue revealed novel viruses associated with nine new archaeal lineages within *Nitrososphaeria*, including *Ca. Caldarchaeales*, *Nitrosocaldales*, *Nitrosotalea* and other AOA-related lineages (also known as non-AO *Thaumarchaeota*), thereby expanding the host spectrum from two to eleven lineages[32-38,49].

Six novel viral families are associated with *Nitrososphaeria*

Nitrososphaeria-associated viruses or MGEs exhibit distinct differences in genome content compared with other viruses. The genome similarity network (sharing at least 20% of protein clusters) yielded 215 groups, with *Nitrososphaeria* contributing to 82 of them (Fig. 2a). These groups included 39 AOA-specific and 24 *Ca. Caldarchaeales*-specific groups, suggesting the distinctiveness of *Nitrososphaeria*-associated viral or MGE genomes within the virosphere. At the genus-to-subfamily viral cluster (VC) level, 95.2% of *Nitrososphaeria*-associated viruses or MGEs formed 119 independent VCs outside of the known ones. However, some *Nitrososphaeria*-associated viruses exhibited close genome-content relationships with other archaeal viruses. Ten shared significant genomic similarity (≥20% shared protein clusters) with established ANME-1-associated viral families (*Chaacviridae*, *Ekchuahviridae* and *Itzamnaviridae*) (Figs. 2a and 2b and Supplementary Dataset 3) [29,30].

Proteome-based hierarchical clustering was employed to further characterize novel families from the *Nitrososphaeria*-associated virome. Twelve viral families were found to be associated with *Nitrososphaeria*, including six previously undescribed ones. Except for the 24 vOTUs that were assigned to established viral families (Fig. 2a and Supplementary Dataset 2), 37 vOTUs were classified into newly named viral families here, they include: one rod-shaped viral family (*Demeterviridae*) and four head-tailed viral families (*Erebusviridae*, *Hypnosviridae*, *Morosviridae* and *Thanatosviridae*), which were associated with *Ca. Caldarchaeales* and a head-tailed family (*Aetherviridae*) associated with AOA (Fig. 3). In addition, 19 MGE families without known capsid proteins were identified, indicating a more complex diversity of the *Nitrososphaeria*-associated virome (Fig. 3 and Supplementary Dataset 2).

The core genomes of different viral families involve ORFs related to replication, structure, packaging, and virus–host recognition. For *Demeterviridae*, in addition to the SIRV2-like capsid, the trimeric autotransporter adhesin (Fig. 3b and Supplementary Dataset 4), one of the key elements of virus–host recognition, was also

part of its core genome. For the other five head-tailed viral families, their core genomes were primarily composed of ORFs associated with essential viral functions, including replication, structure, and packaging. The core genomes of these families tended to be clustered together within the genome, suggesting their conserved functional modules.

The *Nitrososphaeria*-associated viruses exhibit distinct phylogenetic placements within the *Adnaviria* and *Duplodnaviria*

Bayesian and Maximum-Likelihood inference with structural modelling was used to characterize the phylogenetic placements of *Nitrososphaeria*-associated viruses (Extended Data 1 and 2). Two *Caldarchaeales*-associated rod-shaped viruses, together with ANME-1-associated *Ahmunviridae*, formed a phylogenetic clade and identified as a sister group of *Demeterviridae* (Fig. 4a). In the structural similarity network, the capsid monomers and dimers of *Demeterviridae* both formed distinct groups (Figs. 5a and 5b, Extended Data 3). Unlike *Rudiviridae* and *Ahmunviridae*, the major capsid protein (MCP) genes of *Demeterviridae* appear to have undergone duplication events. Although the genomes of *Lipothrixviridae* and *Tristromaviridae* encode two MCP genes and form heterodimeric capsids[50], the MCPs encoded by *Demeterviridae* are more closely related to those of *Rudiviridae* and *Ahmunviridae* (Fig. 4a and Supplementary Figs. 1a and 1b). Considering that *Rudiviridae* and *Ahmunviridae* possess homodimeric capsids, the mode of capsid formation in *Demeterviridae* remains highly uncertain. Nevertheless, the occurrence of multiple copies of MCP genes in *Demeterviridae* represents a rare case within *Adnaviria*. Taken together, these lines of evidence indicate that *Demeterviridae* represent a novel family distinct from other rod-shaped viruses, characterized by unique genome content, phylogenetic placement and structural features of their capsid proteins.

Nitrososphaeria-associated head-tailed viruses exhibit typical polyphyletic characteristics. Most newly identified *Caldarchaeales*-associated viral families formed distinct monophyletic clades in phylogenetic trees of both MCP and terminase,

such as *Morosviridae*, *Erebusviridae* and *Thanatosviridae* (Figs. 4b and 4e). The MCPs encoded by *Morosviridae* form three distinct structurally-similar-based groups (Fig. 5c and Extended Data 4), supporting the uniqueness observed in their phylogenetic inference. These lines of evidence indicate their novel phylogenetic placements within *Caudoviricetes*. Eight *Caldarchaeales*-associated vOTUs were assigned to *Ekchuahviridae*, which had previously been found only in association with ANME-1 [30]. *Ekchuahviridae* was phylogenetically clustered with other archaea viral families (*Ahpuchviridae*, *Halomagnusviridae* and *Haloferuviridae*), suggesting a possible common ancestor of these archaeal head-tailed viruses (Fig. 4c). The majority of AOA-associated viral families belong to established families, such as *Thaspiviridae*, *Yangangviridae* and *Kyanoviridae*. Although *Kyanoviridae* primarily infects cyanobacteria, previous studies have also reported associations with other heterotrophic bacteria, such as *Ca. Pelagibacter*, *Methylophilales* and *Ochrobactrum* [51-53]. Associations between *Kyanoviridae* and AOA were supported by both phylogenetic evidence and the structural similarity of the MCP (Figs. 4d, 4f and 5c), thereby extending the host spectrum of this family to the archaeal domain for the first time. *Aetherviridae* is a newly defined AOA-associated viral family in this work, although members of this family have been described in a previous study [35]. In terminase-based phylogenetic inference, *Aetherviridae* formed a deeply-branched monophyletic clade, indicating its unique phylogenetic placement (Fig. 4e).

Novel auxiliary genes encoded by *Caldarchaeales*-associated rod-shaped viruses

The protein functional diversity of *Nitrososphaeria*-associated viruses or MGEs and their potential impacts on the host were further explored. Following the recommendations of Martin et al., potential auxiliary viral genes (AVGs) encoded by these viruses or MGEs were screened [54]. A total of 288 auxiliary regulatory genes (AReGs), 154 auxiliary metabolic genes (AMGs) and 43 auxiliary physiological genes (APGs) were obtained (Supplementary Dataset 4). For AMGs, 2-polyprenyl-3-methyl-5-hydroxy-6-metoxy-1,4-benzoquinol methylase (UbiG) was

found to be encoded by four vOTUs from the newly identified rod-shaped viral family, *Demeterviridae* (Supplementary Dataset 4). To the best of our knowledge, CoQ is rare in archaea and there is currently no evidence that *Nitrososphaeria* or archaeal viruses possess a complete CoQ biosynthesis pathway [23,55]. Homologues of *Demeterviridae*-encoded UbiG-like protein are extremely rare in viruses, with no high-similarity homologues found in the nr database (alignment identity and coverage both ≥ 50%), highlighting its novelty. These virus-encoded UbiG-like proteins form a distinct monophyletic clade that groups with distantly related archaeal homologues in the phylogenetic tree. This indicates their connection to archaea while retaining a clear virus-specific phylogenetic monophyly (Fig. 6a). This finding provides a rare case of archaeal viruses encoding genes involved in quinone electron carrier biosynthesis [56], suggesting that *Demeterviridae*-encoded UbiG may be involved in modulating quinone electron carriers on the host membrane. For AReGs, *Demeterviridae* was found to encode an ArsR-like transcriptional regulator, which was only previously reported from spindle-shaped ssDNA viruses and head-tailed viruses [49,57]. The *Demeterviridae*-encoded ArsR-like proteins forms a well-supported, long-branched monophyletic clade with a pronounced stem branch in the phylogenetic tree and is evolutionarily distant from other bacteriophages and archaeal viruses. This suggests their unique phylogenetic placement (Fig. 6b). Regulators of archaeal filamentous or rod-shaped viruses are predominantly represented by RHH-fold proteins such as SvtR (*Rudiviridae*) and AvtR (*Lipothrixviridae*) [58-60]. This result suggests that *Caldarchaeales*-associated rod-shaped viral genomes may employ a similar mechanism of transcriptional regulation as spindle-shaped viruses [57], which may represent an unrecognized mechanism within this viral realm.

Viruses or MGEs serve as the primary vectors rather than donors or receptors in horizontal gene transfers

Virus–host linkages identified by spacer–protospacer matching showed lineage specificity, with no cross-lineage protospacer sharing observed (Supplementary Fig.

2a and Supplementary Dataset 5). After establishing virus-host linkages by spacer-protospacer matching, the direction of horizontal gene transfers (HGTs) mediated by viruses or MGEs were explored based on phylogenetic topologies [61]. A total of 2,509 HGT events between viruses or MGEs and corresponding *Nitrososphaeria* lineages were identified (Supplementary Dataset 6 and Extended Data 5) [61]. For AOA, 68.0% of transfers are bidirectional (n = 1,359), while 27.3% originate from archaea as donors (n = 546). Only 4.7% of transfers involve viruses or MGEs as donors (n = 94). Similarly, in *Ca. Caldarchaeales*, 56.5% of transfers are bidirectional (n = 183), 34% originate from archaea (n = 110) and 9.6% involve viruses or MGEs as donors (n = 31) (Fig. 6a). This suggests that viruses or MGEs serve as the primary vectors in these transfers, rather than donors or receptors.

Functionally, most HGTs are related to glycan biosynthesis and metabolism, replication and repair and carbohydrate metabolism, accounting for 60.5% of the total HGTs (Fig. 6b). Within glycan biosynthesis and metabolism, more than half of the HGTs (n = 283) are associated with glycosylation and are linked to Group 1.1a (Supplementary Dataset 4 and Supplementary Dataset 6). The related functional proteins include all dolichyl-phosphate-mannose–protein mannosyltransferases, glycosyl transferases (GTs), sulfatases and N-acetyl sugar amidotransferases. Since glycosylation or glycoprotein modification on the surface of archaeal viral capsids or envelopes is involved in host interactions or enhances structural stability [62-64], this provides evidence for the assimilation of host functional proteins by *Nitrososphaeria*-associated viruses or MGEs. In addition, 87.8% of the HGTs related to viral structure were found to be associated with *Ca. Caldarchaeales*. These structural proteins include tail-related proteins and MCPs from *Erebusviridae* and *Morosviridae*, further confirming the interactions between these proposed viral families and *Ca. Caldarchaeales*.

Discussion

In this study, we provide a deeper understanding of the diversity of

Nitrososphaeria-associated viruses or MGEs, broadening the scope from previously reported AOA-associated viruses and identified 263 previously unrecognized *Caldarchaeales*-associated vOTUs for the first time. This genomic catalogue expands the known number of *Nitrososphaeria*-associated viruses or MGEs by ten-fold [32-36]. The mining of *Caldarchaeales*-associated viral or MGE genomes fills a critical knowledge gap of archaeal viromes. Six previously unknown viral families were identified, including a novel rod-shaped family, *Demeterviridae*. The unique evolutionary placement and functional characteristics of this family were elucidated through comparative genomic and phylogenetic analyses, providing new insights into the archaeal virosphere and viral evolution.

The core genome of *Demeterviridae*, in addition to the SIRV2-like capsid and glycosyl transferase, also contains genes encoding trimeric autotransporter adhesins and several right-handed parallel β-helix repeat proteins (RHBH) (Fig. 3b and Supplementary Dataset 4). Trimeric autotransporter adhesins are key mediators of prokaryote–host recognition [65], but have not previously been reported in viruses. Similarly, RHBH structures are widely found in various enzymes and adhesion factors involved in carbohydrate processing or cell surface recognition [66]. Typically, proteins associated with virus–host recognition are subject to intense selective pressure in the ongoing “arms race” against the frequent alterations of host membrane receptors, and thus are rarely incorporated into the core viral genome [67-69]. The retention of these genes in the core genome of *Demeterviridae* may result from co-evolutionary constraints imposed by other core genes. In other words, these virus–host recognition–related genes may have become integrated into the virion structural and assembly framework [70-72], forming part of a “structural community” whose evolutionary flexibility has been correspondingly reduced.

The phylogenetic analysis of *Nitrososphaeria*-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[60]. Here, it was found that some *Caldarchaeales*-associated rod-shaped viruses encode two copies of MCP genes (Figs. 4a and 5b, Supplementary Figs. 1a and 1b). With

respect to capsid assembly strategy, homodimers may provide symmetry and repetitiveness, thereby contributing to the maintenance of capsid stability [73,74]. Heterodimers may afford greater structural flexibility and diversity, enabling adaptation to varying environmental stresses and host defence mechanisms [75,76]. Given that *Rudiviridae* and *Ahmunviridae* encode only a single-copy MCP gene [30,50,60], it is possible that in *Demeterviridae* only one copy is functional, while the other may have been inactivated under natural selection [77-79]. Nevertheless, the actual capsid assembly mechanism of these rod-shaped viruses remains unknown and these inferences still require direct observation for confirmation. Some *Caldarchaeales*-associated viruses have close relationships with ANME-1-associated viruses or MGEs originating from hydrothermal vents (e.g. Methanophagales virus PBV300, *Ahmunviridae*) (Figs. 2a, 4c and 5b) [30], implying viral host-switching or non-specific transfers occurred between both archaeal lineages, which is common in the evolution of viruses [80-82].

Viruses have been found to encode diverse AVGs that manipulate host metabolism, thereby creating favourable conditions for completing their life cycles and indirectly contributing to global biogeochemical cycles [83-86]. In archaea-associated viruses, such as those infecting methanogenic archaea, Asgard superphylum, halophilic archaea, virus-encoded metabolic genes have been identified that potentially modulate archaeal methane-oxidization, sulfur metabolism, nucleotide and amino acid metabolisms [27,87,88]. Regarding AOA-associated viruses or MGEs, previous studies have discovered that viruses may manipulate host ammonia oxidation metabolism by carrying the *amoC* gene and participate in host vitamin synthesis [34,35]. Through the exploration of functional repertoires for *Nitrososphaeria*-associated viruses, we provide new insights into the functional mechanisms of archaeal viruses. An important example is provided by viral ArsR-like and UbiG-like protein, which represent a rare AReG and AMG, respectively, encoded by a previously unrecognized rod-shaped viral family. If the viral ArsR-like protein is functional in the archaeal virocell, it might be able to hijack the arsenic response of *Ca. Caldarchaeales*, thereby facilitating viral replication within the host cell. Considering that the habitat

of *Ca. Caldarchaeales* is rich in metals and metalloids (including arsenic) and exhibits highly active redox dynamics[24,89,90], this process may be essential for the completion of the *Demeterviridae* life cycle. Phylogenetically, *Demeterviridae*-encoded ArsR falls on a long-branch, virus-specific clade supports their viral origin, rather than recent horizontal gene transferring from archaea.

The more unusual finding is the viral carriage of UbiG-like protein, an enzyme implicated in CoQ biosynthesis. Quinones (e.g. ubiquinone, menaquinone) are key electron carriers in membrane redox chains, yet they are considered rare in archaea and corresponding viruses [23,55,91]. The detection of *Demeterviridae*-encoded UbiG-like protein suggests a potential viral contribution to the redox balance of archaeal membranes, such as altering the local pool of quinone electron carriers [55,92]. They might influence electron flux, respiration efficiency, reactive oxygen species (ROS) levels, or even redox signaling [92-94]. In hyperthermal habitats of *Ca. Caldarchaeales*, control of archaeal quinone pools may confer a fitness advantage to related viruses by maintaining a more favorable intracellular environment for replication. On the other hand, the *Demeterviridae*-encoded UbiG-like protein, together with other methyltransferases encoded by archaeal viruses, may perform similar functions in counteracting host defense mechanisms, representing an exaptation of archaeal gene functions by viruses [81]. Since *Ca. Caldarchaeales* has not been observed to encode UbiG-like proteins, the origin of these genes in *Demeterviridae* remains unknown. Overall, the discovery of *Demeterviridae* not only expands the boundaries of the archaeal virosphere and offers new insights into viral evolution, but also provides direction in how viruses might adapt to extreme environments in terms of transcriptional regulation and host metabolic reprogramming.

Our results reinforce the view that viruses or MGEs act predominantly as gene vectors. This supports the “virus world” model, considering viruses or other selfish genetic elements as intermediaries in gene exchange rather than as main sources of new genes [95]. The prevalence of bidirectional HGTs highlights reciprocal coevolution between viruses or MGEs and archaeal hosts, evoking the arms-race dynamics

observed in archaeal virus–host systems [96,97]. The functional enrichment of glycan biosynthesis and glycosylation among transferred genes implies that viruses may co-opt host surface modification machinery to refine capsid stability or host recognition—a phenomenon similarly observed in eukaryotic viruses adapting envelope glycosylation [98,99]. It is hypothesized that the frequent HGT events between archaea and viruses may have driven the diversification of archaeal envelope biochemistry and the emergence of niche specialization in ammonia-oxidizing lineages.

In summary, the establishment of *Nitrososphaeria*-associated viromic catalogue and our further analysis provides critical knowledge for future in-depth research into virus-archaea interactions.

Methods

Assembly of *Nitrososphaeria*-associated CRISPR database

The *Nitrososphaeria* phylum comprises several major clades: AOA, other AOA-related lineages and *Ca. Caldarchaeales*. To produce a comprehensive genomic dataset, reference genomes were retrieved from GenBank and IMG/M [39,40], 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 [27,28], 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 [100,101]: word size = 7, spacer length = 26 and a minimum of three directed repeats (DRs). This step resulted in the identification of 37,485 *Nitrososphaeria*-associated spacers, which were subsequently used for the detection of viruses or MGEs and downstream analyses.

The metagenome-derived spacers that are associated with *Nitrososphaeria* were

detected to further expand the diversity of these CRISPR spacers. First, a total of 170 metagenome-derived assemblages (including 58 from the Malaspina expedition and 22 from our previous studies) from diverse ecosystems were retrieved and downloaded from IMG/M and ENA (European Nucleotide Archive) (Supplementary Dataset 1)[40,43]. These metagenomes were from 23 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 *Nitrososphaeria* genomes using MMseqs2 with parameters [102]: word size = 7, pairwise alignment identity = 1, percentage coverage of DRs = 1, mismatches = 1. This step produced 196,551 spacers derived from 170 metagenomes. Finally, this process generated a *Nitrososphaeria*-associated spacer database containing 234,036 spacer sequences (Fig. 1a and Extended Data 6).

As in previous studies [28], CRISPR-Cas cassette prediction and subtyping for archaeal contigs and genomes were conducted using CRISPRcasIdentifier v1.0.0 [103], 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 (Supplementary Fig. 2b and Supplementary Dataset 7).

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

The IMG/VR (v.4) database, contigs from 170 metagenomes and archaeal genomes served as the primary resources for mining *Nitrososphaeria*-associated viral or MGE genomes. At first, the archaeal viral or 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 [33-36,38,104,105]. The 37,485 CRISPR spacers from *Nitrososphaeria* 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 or MGEs were also predicted to be *Nitrososphaeria* in IMG/VR (v.4). Due to some viral contigs being assigned to non-*Nitrososphaeria* hosts in IMG/VR (v.4) metadata, the genome-content similarity between viruses or 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 [106]; ii) the *Nitrososphaeria*-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 *Nitrososphaeria* (Supplementary Dataset 2); iv) further validation was conducted to assess the applicability of this screening criterion to previously reported *Nitrososphaeria*-associated viruses or MGEs:

1) the 13 identified *Nitrososphaeria*-associated viral or MGE genomes were retrieved and downloaded from GenBank;
2) the 11 representative genomes were generated from 13 *Nitrososphaeria*-associated viral genomes through MMseqs2 with parameters [102]: 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 [37]. This indicates that the approach is also applicable to most previously reported

Nitrososphaeria-associated viral or 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 *Nitrososphaeria*-associated viruses or MGEs were filtered from 368 spacer-targeted viral or MGE candidates. Similar to previous studies [29,107], all relevant archaeal CRISPR-Cas systems were evaluated using CRISPRCasFinder with default parameters [108]. 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 or 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 or MGEs possessing CRISPR arrays with an evidence level of 4.

In addition, IMG/VR (v.4) contains 1,457 viruses or MGEs that were predicted to be associated with *Nitrososphaeria*, although these contigs were not retrieved through spacer-protospacer matching. The 1,457 viruses or MGEs predicted to be associated with *Nitrososphaeria* 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 *Nitrososphaeria*-associated viruses or MGEs. Finally, 612 *Nitrososphaeria*-associated viruses or MGEs were identified and filtered from IMG/VR (v.4).The completeness of these viral or MGEs were evaluate using CheckV [109].

Expanded mining of *Nitrososphaeria*-associated viruses or MGEs from 170 metagenomes

The *Nitrososphaeria*-associated CRISPR spacer database was aligned to from 170 metagenomes using MMseqs2 with the same approach and parameters [102]. Contigs

containing DRs and those shorter than 2-kb were removed before performing alignment. This resulted in 1,951 contigs matched by *Nitrososphaeria*-associated CRISPR spacers. To remove duplications, 1,951 contigs were aligned to previously identified *Nitrosospaherota*-associated viruses or MGEs set using MMseqs2 with same parameters same parameters [102]. 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 or MGEs, these contigs were used to filter *Nitrososphaeria*-associated viruses or MGEs using three multiple viral identification pipelines, including geNomad, VirSorter2 and DeepVirFinder[110-112]. Multiple pipelines were employed for virus identification based on the following considerations: 1) geNomad demonstrates superior overall classification performance compared to other pipelines [111]; 2) VirSorter2 (VS2) is recommended when integrating multiple pipelines [113]; 3) DeepVirFinder (DVF), utilizing an alignment-free approach, is well-suited for identifying unique viral sequences [110]; 4) the identification results from each pipeline complement one another, enhancing recall while mitigating cellular contamination in the dataset[114-116]. 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 or 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 *Nitrososphaeria*-associated viruses or MGE and were used as a test set to evaluate the efficiency of the approach. Based on the test, this viral or MGE identification approach achieved a 74% recruitment rate, which was considered acceptable for further viral or MGE identification (Supplementary Dataset 8).

Using the validated hybrid viral identification approach, 419 viral or MGE

contigs were identified from 1,396 metagenome-derived contigs. These contigs had already been targeted by *Nitrososphaeria*-associated CRISPR spacers in previous steps. The viral-host genome-content similarity was used to filter these contigs as described before, resulting 108 *Nitrososphaeria*-associated viral or MGEs. Relevant archaeal CRISPR-Cas systems of these contigs were evaluated using CRISPRCasFinder with default parameters [108]. The completeness of these viral or MGEs was evaluated using CheckV [109].

Detection of viruses or MGEs integrated in *Nitrososphaeria* genomes

Proviruses integrated into *Nitrososphaeria* genomes were identified using the "find-proviruses" program in geNomad [111]. Integrases were first detected with the MMseqs2 plugin in geNomad [102], 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 *Nitrososphaeria*-associated proviruses.

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

Host assignments of *Nitrososphaeria*-associated viruses or MGEs

The host assignments of *Nitrososphaeria*-associated viruses or MGEs were primarily based on: 1) CRISPR-spacers-associated host genomes; 2) information provided by

472 previous studies; 3) genome-content similarity between viruses or MGEs and hosts.
473 For *Nitrososphaeria*-associated viruses or MGEs that match CRISPR spacers from
474 multiple distinct hosts, similar with previous study [29], the host of a virus or MGE
475 targeted by CRISPR spacers was assigned based on the taxonomy of the genome from
476 which the spacer with the highest identity originated. For *Nitrososphaeria*-associated
477 viruses or MGEs that were identified based on oligonucleotide pattern similarity, the
478 most possible host genomes associated with these viruses or MGEs were determined
479 through alignment against *Nitrososphaeria* genomes using MMseqs2 with the
480 parameters: 1e-5 as the E-value, iteratively searched two times [102]. A host genome was
481 linked to a viral or MGE genome if they have the most shared ORFs. This approach
482 was also applied in IMG/VR (v.4) [104,117]. For *Nitrososphaeria*-associated viruses or
483 MGEs reported in previous studies, some host information was retrieved from related
484 literature. For those viruses or MGEs with uncertain host information [35], the related
485 viral ORFs were subject to MMseqs2 against all *Nitrososphaeria* genomes with the
486 same approach. The 35 Marthaviruses derived from López-Pérez et al. study were
487 excluded from this study for two reasons [34]: 1) the related data (PRJNA484324) are
488 still embargoed when an attempt was made to access these genomes; 2) although the
489 proposed taxa 'Marthaviruses' has been documented in NCBI, the host information of
490 these viruses is labelled as bacteria. The archaeal host phylogenetic tree was
491 calculated based on 46 concatenated ribosomal proteins by IQ-Tree2 (Supplementary
492 Note 2)[118]. The archaeal lineages were determined based on their phylogenetic
493 placements and previous studies [2,8,21,22,24,119].

494 **Comparative genomics and family-/species-level demarcation of**
495 ***Nitrososphaeria*-associated viruses or MGEs**

496 Similar to previous studies [27,29,31], new families and vOTUs were assigned based on
497 the genome-content similarity network and the genome-wide proteomic tree. Using
498 DNA virus sequences longer than 2-kb from NCBI RefSeq v.223 as reference
499 genomes, vConTACT2 was applied (with default parameters) to compute the

genome-content-based similarity network of those *Nitrososphaeria*-associated viruses or MGEs. *Nitrososphaeria*-associated viral or 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 *Nitrososphaeria*-associated viruses or 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 or 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) [29]. The network and tree were visualized by Cytoscape v.3.10.2 and iToL v.6, respectively [120,121]. The nomenclature of the proposed viral family was referred to Cthulhu mythos (Supplementary Note 3).

Annotation of viral or MGEs genomes

Viral or MGE species (vOTUs or MGE OTUs) were defined using 95% nucleotide identity threshold over 90% of the full length using MMseqs2. The family-level taxonomic assignments of *Nitrososphaeria*-associated viruses or 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 [104,122,123]; 2) CAT with NCBI RefSeq v.223 database [123,124]; 3) geNomad with default parameters [111]. 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 [125]. The members of each proposed family were furtherly expanded based on consensus vOTU.

ORFs encoded in *Nitrososphaeria*-associated viruses or MGEs were predicted by Prodigal v.2.6.3 with metagenome mode[106] and annotated using the KEGG Ortholog database (January, 2025), Pfam-A v.35, COG2020, Swiss-Prot database (January, 2025), VOG223 database and NR database (August, 2025) [126-128]. Potential AVGs encoded by these viruses or MGEs with length over 10-kb were screened based on proposal from Martin et al.[54]. The functional screening criteria excluded all genes associated with viral core functions (replication, recombination, repair, transcription, translation, packaging and lysis), nucleotide metabolism and one-carbon pool processes, nucleic acid modification and immune evasion, structural and host recognition functions (including all CAZymes and polysaccharide synthesis related fucntions), dNTP supply, as well as various proteases or multifunctional enzymes. In addition, candidate genes were filtered by examining whether viral genes were present in their flanking regions. Considering the incompleteness of viral contigs, once an AVG was confirmed, its homologues were also considered as AVGs, even if no viral genes were present in the flanking regions. The maximum-likelihood phylogenetic trees of ArsR and UbiG were constructed. For ArsR, protein sequences from archaea, bacteria, bacteriophages and archaeal viruses were retrieved from InterPro (PF01022). After merging these with the ArsR protein sequences encoded by rod-shaped viruses, CD-HIT v. 4.8 was used to cluster the sequences from different sources (at 0.5 pairwise identity with global alignment). Seqkit v.2.5 was then applied to randomly sample 100 sequences from each category [129], or to include all sequences if fewer than 100 were available. Multiple sequence alignment (MSA) was performed with MUSCLE5 [130] and trimAl v.1.4 was used to remove sites with more than 30% gaps [131]. Maximum-likelihood phylogenetic inference of the MSAs was carried out with IQ-TREE2 [118], with 1,000 bootstraps and LG+F+R10 as the best-fit substitution

model (Extended Data 7). For UbiG, protein sequences from archaea, bacteria and eukaryotes were retrieved from InterPro (IPR004951). These were merged with rod-shaped virus-encoded UbiG-like sequences and maximum-likelihood phylogenetic inference was performed using a similar procedure, with Blosum62+R10 as the best-fit substitution model (Extended Data 7). The maximum-likelihood phylogenetic trees were visualized with iTOL v.6 [121].

Determining of *Nitrososphaeria*-associated viral or MGE horizontal gene transfer events

HGT events between *Nitrososphaeria*-associated viruses (vOTUs) and their archaeal hosts were identified following an established pipeline [61]. 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 9).

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 *Nitrososphaeria*-associated viruses or MGEs

To investigate the evolutionary relationships among rod-shaped and head-tailed viruses associated with *Nitrososphaeria* within the virosphere context, phylogenetic analyses were performed based on key viral proteins, including the SIRV2-like MCPs,

HK97-like MCPs and terminase proteins. For rod-shaped viruses, MCP sequences from 33 viruses or MGEs were aligned with 34 viral RefSeqs v.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 [132], 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 [127,133]; 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 [118,134], 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 [135].

Structure analysis of *Nitrososphaeria*-associated viruses or MGEs

To further assess the novelty of capsid proteins from *Nitrososphaeria*-associated viruses or 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 [106]. 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 [136], 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 [102], yielding 468 representative sequences. Structures of the HK97-like and SIRV2-like capsid proteins were predicted using AlphaFold2 implemented in ColabFold (Extended Data 3 and 4) [137] 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 [132] 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[120,132].

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. Abby, S.S., Kerou, M., and Schleper, C. (2020). Ancestral Reconstructions Decipher Major Adaptations of Ammonia-Oxidizing Archaea upon Radiation into Moderate Terrestrial and Marine Environments. mBio *11*. 10.1128/mBio.02371-20.
17. 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.
18. 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.
19. 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.
20. 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.
21. 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.
22. 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.
23. 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.
24. 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.
25. 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.
26. Al-Shayeb, B., Schoelmerich, M.C., West-Roberts, J., Valentin-Alvarado, L.E., Sachdeva, R., Mullen, S., Crits-Christoph, A., Wilkins, M.J., Williams, K.H., Doudna, J.A., and Banfield, J.F. (2022). Borgs are giant genetic elements with potential to expand metabolic capacity. Nature *610*, 731-736. 10.1038/s41586-022-05256-1.
27. 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.
28. 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.
29. 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.
30. 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.
31. 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.

32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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.
41. 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.
42. 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.
43. 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.
44. Gao, C., Liang, Y., Jiang, Y., Paez-Espino, D., Han, M., Gu, C., Wang, M., Yang, Y., Liu, F., Yang, Q., et al. (2022). Virioplankton assemblages from challenger deep, the deepest place in the oceans. iScience *25*, 104680. 10.1016/j.isci.2022.104680.
45. 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.
46. 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.

47. 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.
48. 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.
49. 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.
50. Goulet, A., Blangy, S., Redder, P., Prangishvili, D., Felisberto-Rodrigues, C., Forterre, P., Campanacci, V., and Cambillau, C. (2009). Acidianus filamentous virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage. Proc Natl Acad Sci U S A *106*, 21155-21160. 10.1073/pnas.0909893106.
51. Zhao, Y., Temperton, B., Thrash, J.C., Schwalbach, M.S., Vergin, K.L., Landry, Z.C., Ellisman, M., Deerinck, T., Sullivan, M.B., and Giovannoni, S.J. (2013). Abundant SAR11 viruses in the ocean. Nature *494*, 357-360. 10.1038/nature11921.
52. Buchholz, H.H., Bolanos, L.M., Bell, A.G., Michelsen, M.L., Allen, M.J., and Temperton, B. (2022). A Novel and Ubiquitous Marine Methylophage Provides Insights into Viral-Host Coevolution and Possible Host-Range Expansion in Streamlined Marine Heterotrophic Bacteria. Appl Environ Microbiol *88*, e0025522. 10.1128/aem.00255-22.
53. Decewicz, P., Golec, P., Szymczak, M., Radlinska, M., and Dziewit, L. (2020). Identification and Characterization of the First Virulent Phages, Including a Novel Jumbo Virus, Infecting Ochrobactrum spp. Int J Mol Sci *21*. 10.3390/ijms21062096.
54. Martin, C., Emerson, J.B., Roux, S., and Anantharaman, K. (2025). A call for caution in the biological interpretation of viral auxiliary metabolic genes. Nat Microbiol *10*, 2122-2129. 10.1038/s41564-025-02095-4.
55. Elling, F.J., Becker, K.W., Konneke, M., Schroder, J.M., Kellermann, M.Y., Thomm, M., and Hinrichs, K.U. (2016). Respiratory quinones in Archaea: phylogenetic distribution and application as biomarkers in the marine environment. Environ Microbiol *18*, 692-707. 10.1111/1462-2920.13086.
56. Guerra, R.M., and Pagliarini, D.J. (2023). Coenzyme Q biochemistry and biosynthesis. Trends Biochem Sci *48*, 463-476. 10.1016/j.tibs.2022.12.006.
57. Frols, S., Gordon, P.M., Panlilio, M.A., Schleper, C., and Sensen, C.W. (2007). Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays. Virology *365*, 48-59. 10.1016/j.virol.2007.03.033.
58. Guilliere, F., Peixeiro, N., Kessler, A., Raynal, B., Desnoues, N., Keller, J., Delepierre, M., Prangishvili, D., Sezonov, G., and Guijarro, J.I. (2009). Structure, function, and targets of the transcriptional regulator SvtR from the hyperthermophilic archaeal virus SIRV1. J Biol Chem *284*, 22222-22237. 10.1074/jbc.M109.029850.
59. Peixeiro, N., Keller, J., Collinet, B., Leulliot, N., Campanacci, V., Cortez, D., Cambillau, C., Nitta, K.R., Vincentelli, R., Forterre, P., et al. (2013). Structure and function of AvtR, a novel transcriptional regulator from a hyperthermophilic archaeal lipothrixvirus. J Virol *87*, 124-136. 10.1128/JVI.01306-12.
60. 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.
61. 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.

62. Gebhard, L.J., Vershinin, Z., Alarcon-Schumacher, T., Eichler, J., and Erdmann, S. (2023). Influence of N-Glycosylation on Virus-Host Interactions in Halorubrum lacusprofundi. Viruses *15*. 10.3390/v15071469.
63. Quemin, E.R., Pietila, M.K., Oksanen, H.M., Forterre, P., Rijpstra, W.I., Schouten, S., Bamford, D.H., Prangishvili, D., and Krupovic, M. (2015). Sulfolobus Spindle-Shaped Virus 1 Contains Glycosylated Capsid Proteins, a Cellular Chromatin Protein, and Host-Derived Lipids. J Virol *89*, 11681-11691. 10.1128/JVI.02270-15.
64. Wang, F., Cvirkaite-Krupovic, V., Kreutzberger, M.A.B., Su, Z., de Oliveira, G.A.P., Osinski, T., Sherman, N., DiMaio, F., Wall, J.S., Prangishvili, D., et al. (2019). An extensively glycosylated archaeal pilus survives extreme conditions. Nat Microbiol *4*, 1401-1410. 10.1038/s41564-019-0458-x.
65. Kiessling, A.R., Malik, A., and Goldman, A. (2020). Recent advances in the understanding of trimeric autotransporter adhesins. Med Microbiol Immunol *209*, 233-242. 10.1007/s00430-019-00652-3.
66. Burnim, A.A., Dufault-Thompson, K., and Jiang, X. (2024). The three-sided right-handed beta-helix is a versatile fold for glycan interactions. Glycobiology *34*. 10.1093/glycob/cwae037.
67. de Leeuw, M., Baron, M., Ben David, O., and Kushmaro, A. (2020). Molecular Insights into Bacteriophage Evolution toward Its Host. Viruses *12*. 10.3390/v12101132.
68. Yehl, K., Lemire, S., Yang, A.C., Ando, H., Mimee, M., Torres, M.T., de la Fuente-Nunez, C., and Lu, T.K. (2019). Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. Cell *179*, 459-469 e459. 10.1016/j.cell.2019.09.015.
69. Taslem Mourosi, J., Awe, A., Guo, W., Batra, H., Ganesh, H., Wu, X., and Zhu, J. (2022). Understanding Bacteriophage Tail Fiber Interaction with Host Surface Receptor: The Key "Blueprint" for Reprogramming Phage Host Range. Int J Mol Sci *23*. 10.3390/ijms232012146.
70. Zheng, K., Liang, Y., Paez-Espino, D., Zou, X., Gao, C., Shao, H., Sung, Y.Y., Mok, W.J., Wong, L.L., Zhang, Y.Z., et al. (2023). Identification of hidden N4-like viruses and their interactions with hosts. mSystems *8*, e0019723. 10.1128/msystems.00197-23.
71. Champeimont, R., Laine, E., Hu, S.W., Penin, F., and Carbone, A. (2016). Coevolution analysis of Hepatitis C virus genome to identify the structural and functional dependency network of viral proteins. Sci Rep *6*, 26401. 10.1038/srep26401.
72. Krupovic, M., and Koonin, E.V. (2017). Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A *114*, E2401-E2410. 10.1073/pnas.1621061114.
73. 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.
74. 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.
75. Mueller, K.L. (2016). Structural insights into capsid flexibility. Science *354*, 1387-1389. 10.1126/science.354.6318.1387-p.
76. 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.
77. Brennan, G., Stoian, A.M.M., Yu, H., Rahman, M.J., Banerjee, S., Stroup, J.N., Park, C., Tazi, L., and Rothenburg, S. (2023). Molecular Mechanisms of Poxvirus Evolution. mBio *14*, e0152622. 10.1128/mbio.01526-22.
78. Monzon, S., Varona, S., Negredo, A., Vidal-Freire, S., Patino-Galindo, J.A., Ferressini-Gerpe, N., Zaballos, A., Orviz, E., Ayerdi, O., Munoz-Gomez, A., et al. (2024). Monkeypox virus genomic accordion strategies. Nat

Commun *15*, 3059. 10.1038/s41467-024-46949-7.

79. Pope, W.H., Jacobs-Sera, D., Russell, D.A., Peebles, C.L., Al-Atrache, Z., Alcoser, T.A., Alexander, L.M., Alfano, M.B., Alford, S.T., Amy, N.E., et al. (2011). Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. PLoS One *6*, e16329. 10.1371/journal.pone.0016329.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. Colman, D.R., Keller, L.M., Arteaga-Pozo, E., Andrade-Barahona, E., St Clair, B., Shoemaker, A., Cox, A., and Boyd, E.S. (2024). Covariation of hot spring geochemistry with microbial genomic diversity, function, and evolution. Nat Commun *15*, 7506. 10.1038/s41467-024-51841-5.
90. Yin, Z., Ye, L., and Jing, C. (2022). Genome-Resolved Metagenomics and Metatranscriptomics Reveal that Aquificae Dominates Arsenate Reduction in Tengchong Geothermal Springs. Environ Sci Technol *56*, 16473-16482. 10.1021/acs.est.2c05764.
91. Yuan, L., and Ju, F. (2023). Potential Auxiliary Metabolic Capabilities and Activities Reveal Biochemical Impacts of Viruses in Municipal Wastewater Treatment Plants. Environ Sci Technol *57*, 5485-5498. 10.1021/acs.est.2c07800.
92. Turunen, M., Olsson, J., and Dallner, G. (2004). Metabolism and function of coenzyme Q. Biochim Biophys Acta *1660*, 171-199. 10.1016/j.bbamem.2003.11.012.
93. Bekker, M., Alexeeva, S., Laan, W., Sawers, G., Teixeira de Mattos, J., and Hellingwerf, K. (2010). The ArcBA two-component system of Escherichia coli is regulated by the redox state of both the ubiquinone and the menaquinone pool. J Bacteriol *192*, 746-754. 10.1128/JB.01156-09.
94. Sarewicz, M., and Osyczka, A. (2015). Electronic connection between the quinone and cytochrome C redox pools and its role in regulation of mitochondrial electron transport and redox signaling. Physiol Rev *95*, 219-243.

10.1152/physrev.00006.2014.

95. Koonin, E.V., Dolja, V.V., Krupovic, M., and Kuhn, J.H. (2021). Viruses Defined by the Position of the Virosphere within the Replicator Space. Microbiol Mol Biol Rev *85*, e0019320. 10.1128/MMBR.00193-20.

96. Alvarez, L.M., and Peng, X. (2025). Redefining paradigms in the archaeal virus-host arms race. Preprint. 10.1101/2025.04.20.649705.

97. Lopez-Beltran, A., Botelho, J., and Iranzo, J. (2024). Dynamics of CRISPR-mediated virus-host interactions in the human gut microbiome. ISME J *18*. 10.1093/ismejo/wrae134.

98. Koonin, E.V., Dolja, V.V., and Krupovic, M. (2021). The healthy human virome: from virus-host symbiosis to disease. Curr Opin Virol *47*, 86-94. 10.1016/j.coviro.2021.02.002.

99. Koonin, E.V., Dolja, V.V., and Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology *479-480*, 2-25. 10.1016/j.virol.2015.02.039.

100. 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.

101. 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.

102. 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.

103. 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.

104. 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.

105. 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.

106. 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.

107. 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.

108. 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.

109. 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.

110. 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.

111. 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.

112. 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.

113. 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.

114. 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.

115. 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.

116. 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.

117. 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.

118. 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.

119. 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.

120. 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.

121. 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.

122. 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.

123. 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.

124. 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.

125. 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.

126. 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.

127. 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.

128. 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.

129. Shen, W., Sipos, B., and Zhao, L. (2024). SeqKit2: A Swiss army knife for sequence and alignment processing. Imeta *3*, e191. 10.1002/imt2.191.

130. Edgar, R.C. (2022). Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun *13*, 6968. 10.1038/s41467-022-34630-w.

131. Capella-Gutierrez, S., Silla-Martinez, J.M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972-1973. 10.1093/bioinformatics/btp348.

132. 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.

133. 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.

134. 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.

135. 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.

136. 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.

137. 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.

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 “Expanding diversity of virome associated with ammonium-oxidizing archaea and related lineages”.

Figure legends

Figure 1. The expanding diversity of *Nitrososphaeria*-associated viruses or mobile genetic elements (MGEs). a, The workflow for mining, filtering, and identifying viruses/MGEs primarily included the prediction of archaeal CRISPR systems, viruses or MGEs mining, genome filtering, and the definition of viral operational taxonomic units (vOTUs or MGE OTUs). Only genomes encoding detectable capsid or related proteins were considered viruses; others were classified as MGEs. b, Biogeographic distribution of Nitrososphaerota-associated viral or MGE genomes. In the map, nodes of different colors indicate their putative archaeal hosts: *Ca. Caldarchaeales*, ammonia-oxidizing archaea (AOA), deeply-branched AOA-related archaea (DB-AO-related), and deeply-branched *Nitrososphaeria* archaea (D-B *Nitrososphaeria*). c, The number of vOTUs or MGE OTUs with distinct putative morphologies were originated from different habitats and were associated with diverse *Nitrososphaeria* lineages. Bar colours represent different archaeal lineages. d, Phylogenetic tree of *Nitrososphaeria* genomes. This tree was constructed based on concatenated alignments of 46 ribosomal proteins (totalling 5,963 amino acid positions) using the maximum likelihood method with the LG+C60+F model as recommended by Spang et al. (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 or MGEs with different possible morphologies (at the OTU level) associated with each archaeal lineage.

Figure 2. The similarity of *Nitrososphaeria*-associated viruses or mobile genetic elements (MGEs) based on their genome contents. a, Genome-content similarity network of *Nitrososphaeria*-associated viruses or MGEs within the context of the global virosphere (sequences from NCBI RefSeq v. 223). For clarity, only cluster communities involving *Nitrososphaeria*-associated viruses or MGEs are shown. Each node represents a viral or MGE contig and is coloured according to its host lineage. *Nitrososphaeria*-associated viruses or MGEs clustering with established viral families are annotated adjacent to their respective groups. The pie charts illustrate the number of *Nitrososphaeria*-specific viral clusters (VCs) generated from this network. b, Similarity network of *Nitrososphaeria*-associated viruses or MGEs. Nodes represent viral or MGE genomes and are coloured according to their associated archaeal lineages. Each distinct super viral cluster (sVC) may consist of one or more 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

or MGEs.

Figure 3. Assignments of new families within *Nitrososphaeria*-associated viruses or mobile genetic elements (MGEs). a, Viral proteomic hierarchy clustering 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 (see Methods). The names of the proposed 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-AOA-related, deeply-branched AOA-related archaea; D-B *Nitrososphaeria*, deeply-branched *Nitrososphaeria* archaea. b, Genome organisation diagrams representing each new viral family. Arrows denote individual open reading frames (ORFs), with the colour gradient indicating the sharing rate of each ORF among members within the family. ORFs with a sharing rate greater than 0.9 are considered to encode viral hallmark proteins (VHPs). Asterisks of different colours below the corresponding ORFs indicate the functional categories of the VHPs.

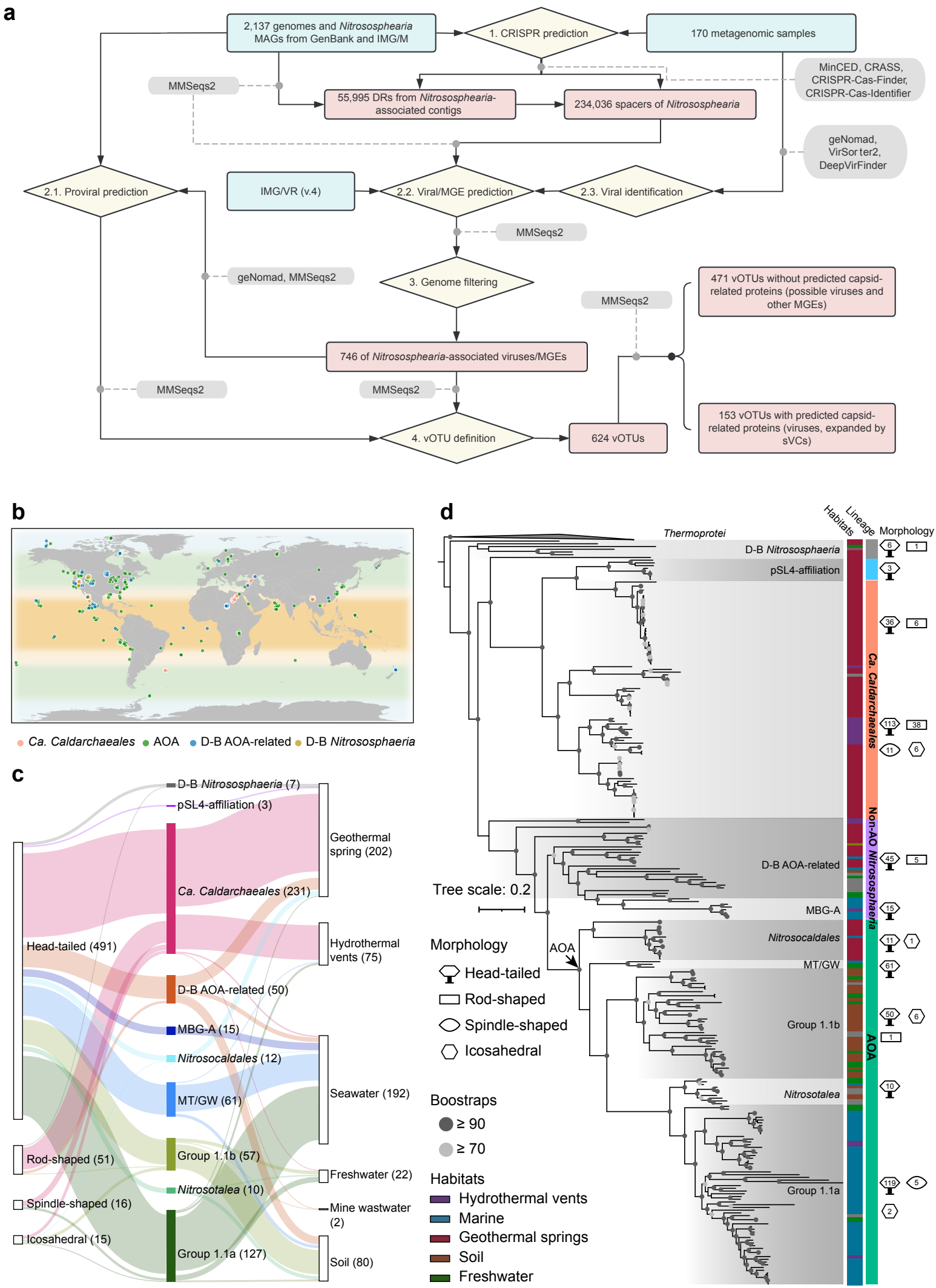
Figure 4. Phylogenetic inferences of *Nitrososphaeria*-associated viruses or mobile genetic elements (MGEs) based on viral hallmark proteins (VHPs) through Bayesian inference (BI) and Maximum Likelihood (ML) approach. a, The trees display the topologies inferred by BI based on VHPs. *Nitrososphaeria*-associated viruses or 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 *Nitrososphaeria* lineages associated with these viruses/MGEs. a, BI 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, BI 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, BI tree based on the HK97-like capsid protein (domain accession in VOGdb: VOG01305), rooted using C1_1 (domain accession in Pfam:

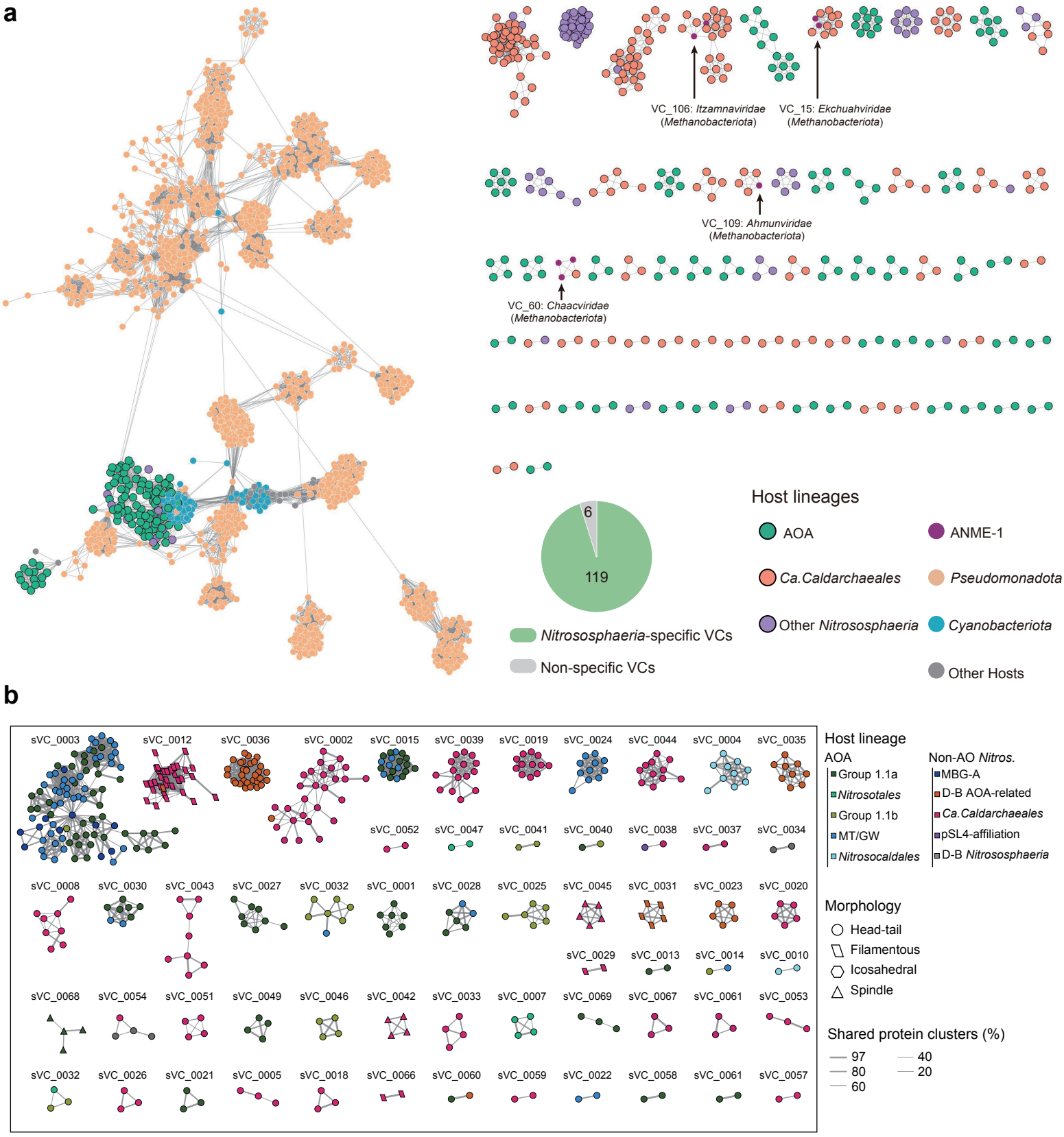
PF00130) as the outgroup. d, BI tree based on the HK97-like capsid protein (domain accession in Pfam: PF07068), rooted using L_lactis_ph-MCP (domain accession in Pfam: PF06673) as the outgroup. e, BI tree based on the terminase large subunit protein (domain accession in KEGG: K06909), rooted using P-loop NTPase (domain accession in SUPERFAMILY: SSF52540) as the outgroup. f, BI 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-AOA-related, deeply-branched AOA-related archaea; D-B *Nitrososphaeria*, deeply-branched *Nitrososphaeria* archaea.

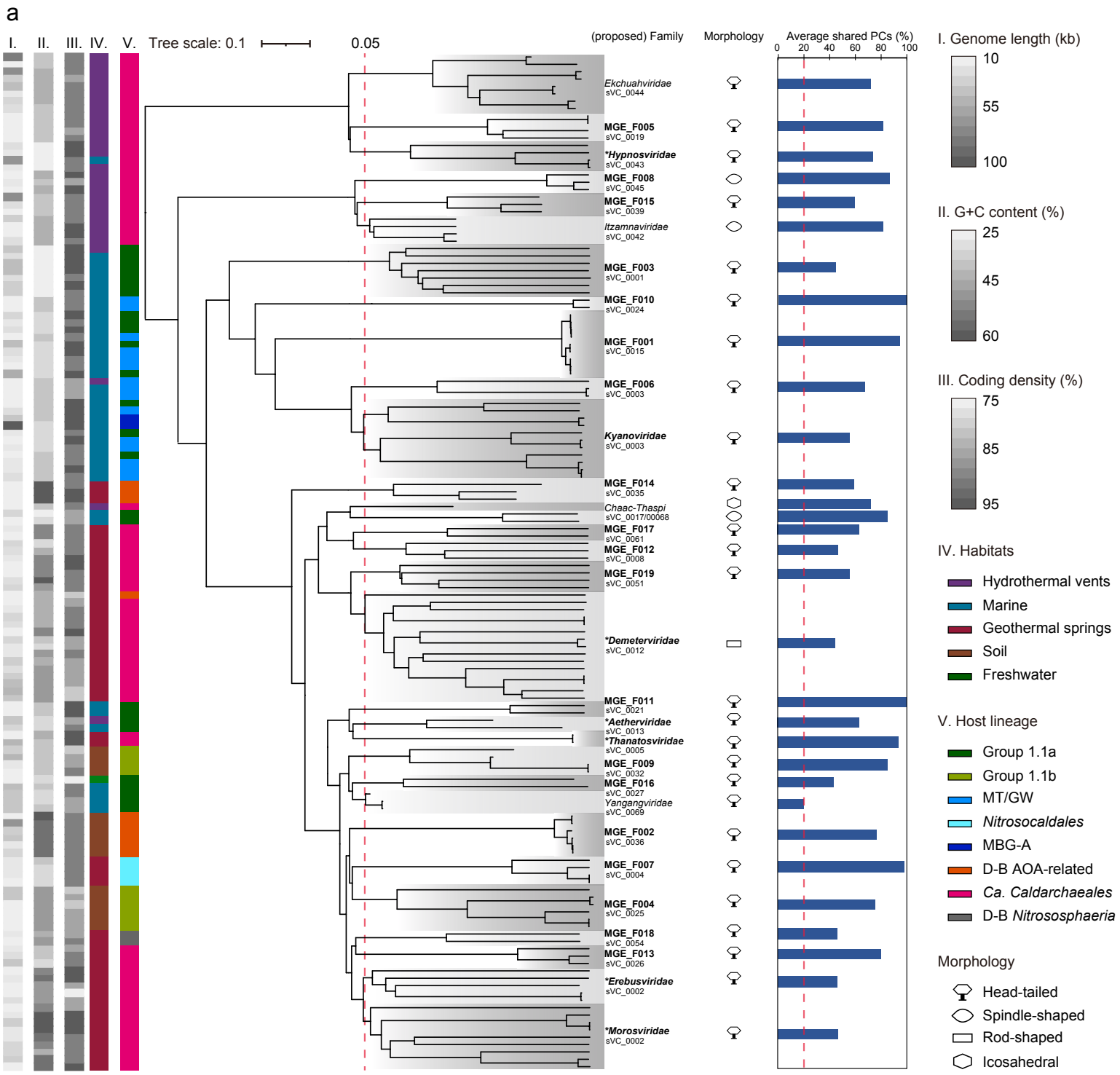
Figure 5. Capsid structure features of *Nitrososphaeria*-associated rod-shaped and head-tailed viruses. a, Three-dimensional structural prediction of archaeal rod-shaped viral capsids and their structure-based similarity network. Nodes represent viral capsids and are color-coded according to their viral families. Edge width indicates the degree of structural similarity between capsids (only identities ≥ 0.5 are shown). Representative structures from each cluster are displayed adjacent to their respective groups. a, predicted structures of viral capsid protein monomers and their clustering network; b, predicted structures of viral capsid protein dimers and their clustering network. Possible heterodimers are indicated with black arrows. Genome IDs and corresponding per-residue measures of local confidence (pLDDT) are indicated below each structure. c, Three-dimensional structural prediction of archaeal HK97-like viral capsids and their structure-based similarity network.

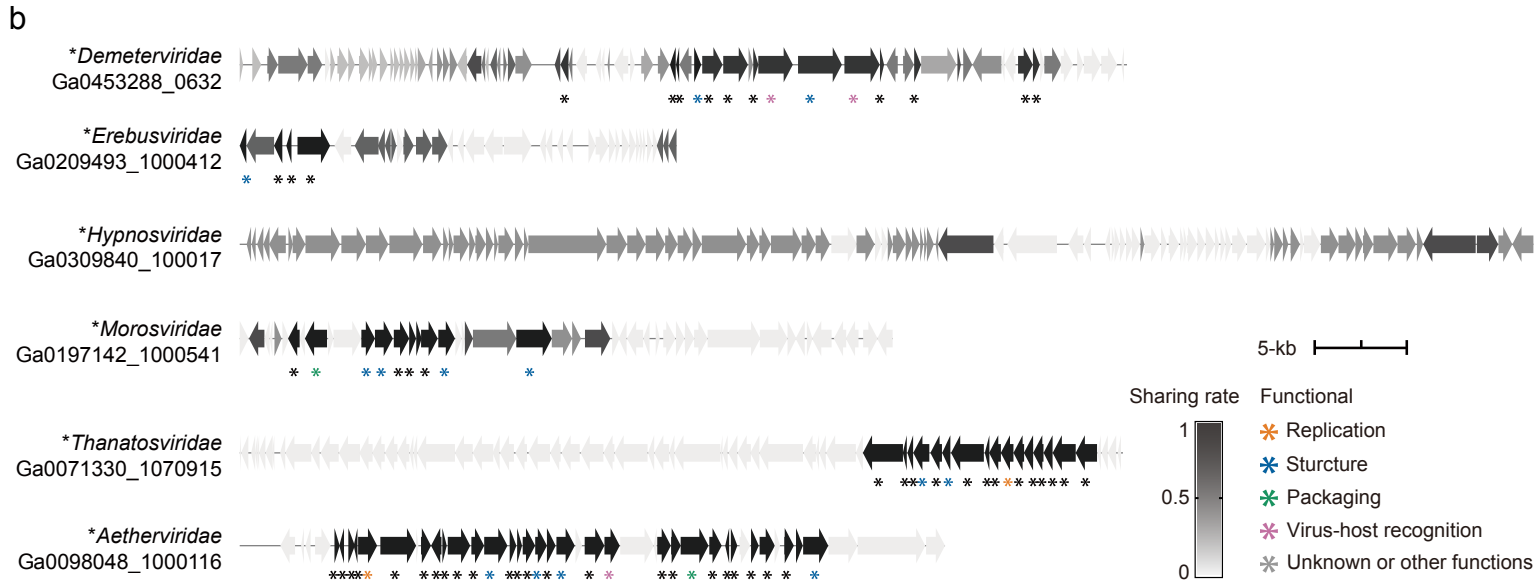
Figure 6. *Nitrososphaeria*-associated viruses encode unique auxiliary genes and exhibit a distinctive landscape of horizontal gene transfer (HGT) with their hosts. a, Maximum-likelihood (ML) phylogenetic tree of the *Demeterviridae*-encoded ArsR-like transcriptional regulator. Branches associated with archaeal viruses and bacteriophages are highlighted in purple and black, respectively, and are shown in bold. The names of archaea-associated viral families are annotated alongside the corresponding branches on the tree. The tree annotation includes proteins of cellular origin (from InterPro), derived from archaeal or bacterial genomes. For clarity, only clades or branches related to viruses were retained, while the others were pruned. Representative structures and their per-residue measures of local confidence (pLDDT) are displayed below the phylogenetic tree. b, Maximum-likelihood (ML) phylogenetic tree of the *Demeterviridae*-encoded UbiG-like protein. c, Proportion of HGT events with different transfer directions among various *Nitrososphaeria* lineages. Different colours of the sectors indicate the direction of HGT, and the numbers within them represent the number of observations. c, c, The number of HGTs with distinct directions were associated from different

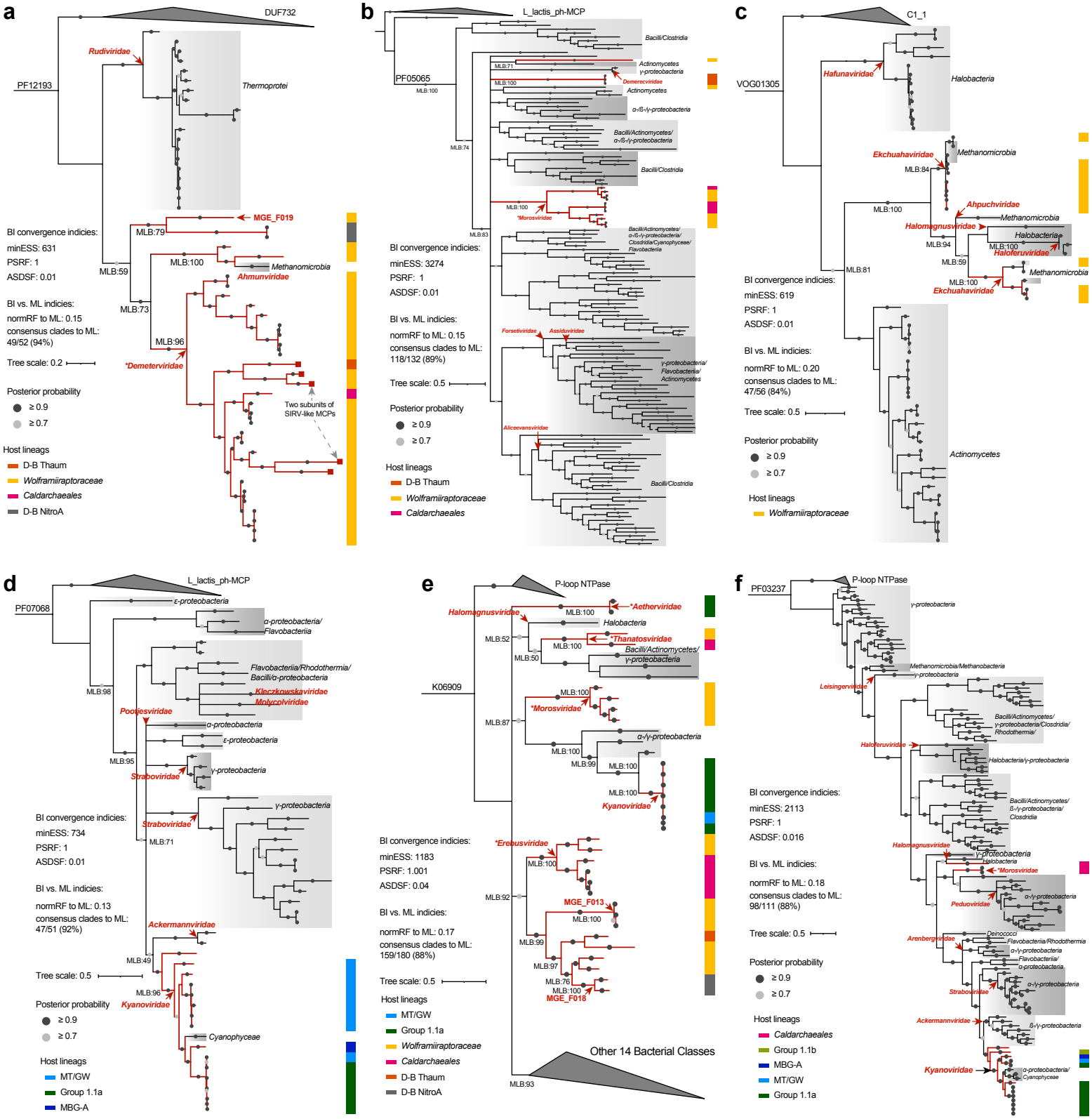
functional catagories and were linked with diverse *Nitrososphaeria* lineages. Bar colours represent the direction of HGT and *Nitrososphaeria* lineages, respectively. The numbers of observations were indicated beside bars. Abbreviations: HTV27, Halorubrum tailed virus 27; SSV5, Sulfolobus spindle-shaped virus 5; phiCDIF1296T, Peptoclostridium phage phiCDIF1296T; MT/GW, Marine Trench and Ground Water; MBG-A, Marine Beneath Group A; D-B-AOA-related, deeply-branched AOA-related archaea; D-B *Nitrososphaeria*, deeply-branched *Nitrososphaeria* archaea.

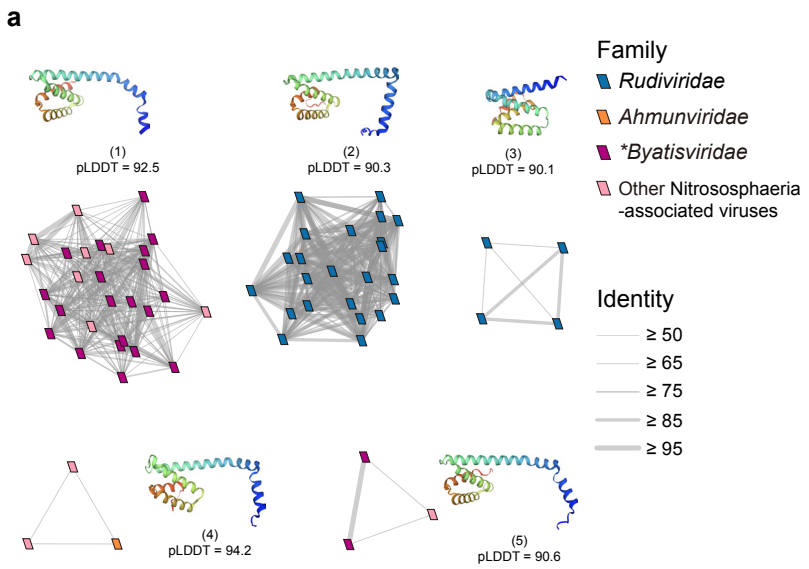

a
2,137 genomes and Nitrososphearia MAGs from GenBank and IMG/M
1. CRISPR prediction
170 metagenomic samples
MMSeqs2
55,995 DRs from Nitrososphearia-associated contigs
234,036 spacers of Nitrososphearia
MinCED, CRASS, CRISPR-Cas-Finder, CRISPR-Cas-Identifier
geNomad, VirSorter2, DeepVirFinder
2.1. Proviral 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 Nitrososphearia-associated viruses/MGEs
MMSeqs2
MMSeqs2
153 vOTUs with predicted capsid-related proteins (viruses, expanded by sVCs)
4. vOTU definition
624 vOTUs
b
Ca. Caldarchaeales
AOA
D-B AOA-related
D-B Nitrososphaeria
c
D-B Nitrososphaeria (7)
pSL4-affiliation (3)
Ca. Caldarchaeales (231)
D-B AOA-related (50)
MBG-A (15)
Nitrosocaldales (12)
MT/GW (61)
Group 1.1b (57)
Nitrosotalea (10)
Group 1.1a (127)
Head-tailed (491)
Rod-shaped (51)
Spindle-shaped (16)
Icosahedral (15)
Geothermal spring (202)
Hydrothermal vents (75)
Seawater (192)
Freshwater (22)
Mine wastwater (2)
Soil (80)
d
Habitats
Lineage
Morphology
Thermoprotei
D-B Nitrososphaeria
pSL4-affiliation
Ca. Caldarchaeales
Non-AO Nitrososphaeria
AOA
D-B AOA-related
MBG-A
Nitrosocaldales
MT/GW
Group 1.1b
Nitrosotalea
Group 1.1a
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
VC_106: Itzamnaviridae (Methanobacteriota)
VC_15: Ekchuahviridae (Methanobacteriota)
VC_109: Ahmunviridae (Methanobacteriota)
VC_60: Chaacviridae (Methanobacteriota)
6
119
Nitrososphaeria-specific VCs
Non-specific VCs
Host lineages
AOA
ANME-1
Ca.Caldarchaeales
Pseudomonadota
Other Nitrososphaeria
Cyanobacteriota
Other Hosts
b
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
AOA
Group 1.1a
Nitrosotales
Group 1.1b
MT/GW
Nitrosocaldales
Non-AO Nitros.
MBG-A
D-B AOA-related
Ca.Caldarchaeales
pSL4-affiliation
D-B Nitrososphaeria
Morphology
Head-tail
Filamentous
Icosahedral
Spindle
Shared protein clusters (%)
97
80
60
40
20

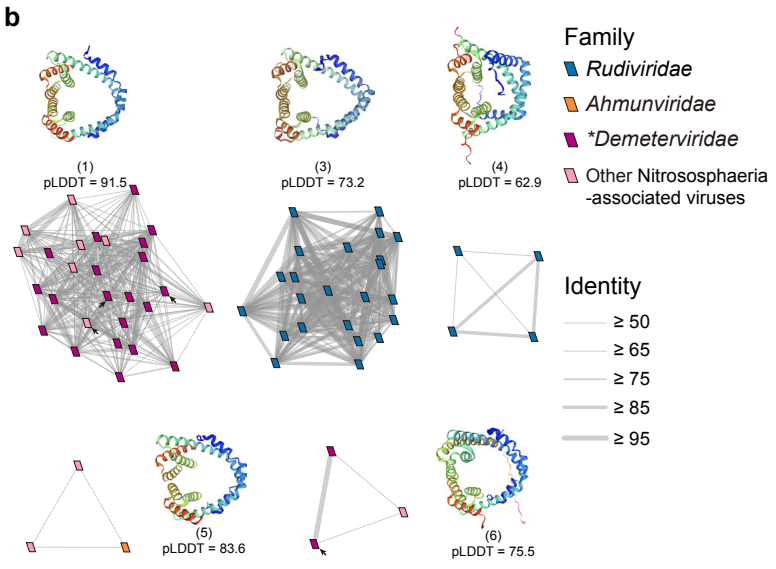

a
I.
II.
III.
IV.
V.
Tree scale: 0.1
0.05
(proposed) Family
Morphology
Average shared PCs (%)
0
20
40
60
80
100
Ekchuahviridae
sVC_0044
MGE_F005
sVC_0019
*Hypnosviridae
sVC_0043
MGE_F008
sVC_0045
MGE_F015
sVC_0039
Itzamnaviridae
sVC_0042
MGE_F003
sVC_0001
MGE_F010
sVC_0024
MGE_F001
sVC_0015
MGE_F006
sVC_0003
Kyanoviridae
sVC_0003
MGE_F014
sVC_0035
Chaac-Thaspi
sVC_0017/00068
MGE_F017
sVC_0061
MGE_F012
sVC_0008
MGE_F019
sVC_0051
*Demeterviridae
sVC_0012
MGE_F011
sVC_0021
*Aetherviridae
sVC_0013
*Thanatosviridae
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
*Erebusviridae
sVC_0002
*Morosviridae
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 AOA-related
Ca. Caldarchaeales
D-B Nitrososphaeria
Morphology
Head-tailed
Spindle-shaped
Rod-shaped
Icosahedral


b
*Demeterviridae
Ga0453288_0632
*Erebusviridae
Ga0209493_1000412
*Hypnosviridae
Ga0309840_100017
*Morosviridae
Ga0197142_1000541
*Thanatosviridae
Ga0071330_1070915
*Aetherviridae
Ga0098048_1000116
5-kb
Sharing rate
1
0.5
0
Functional
Replication
Sturcture
Packaging
Virus-host recognition
Unknown or other functions

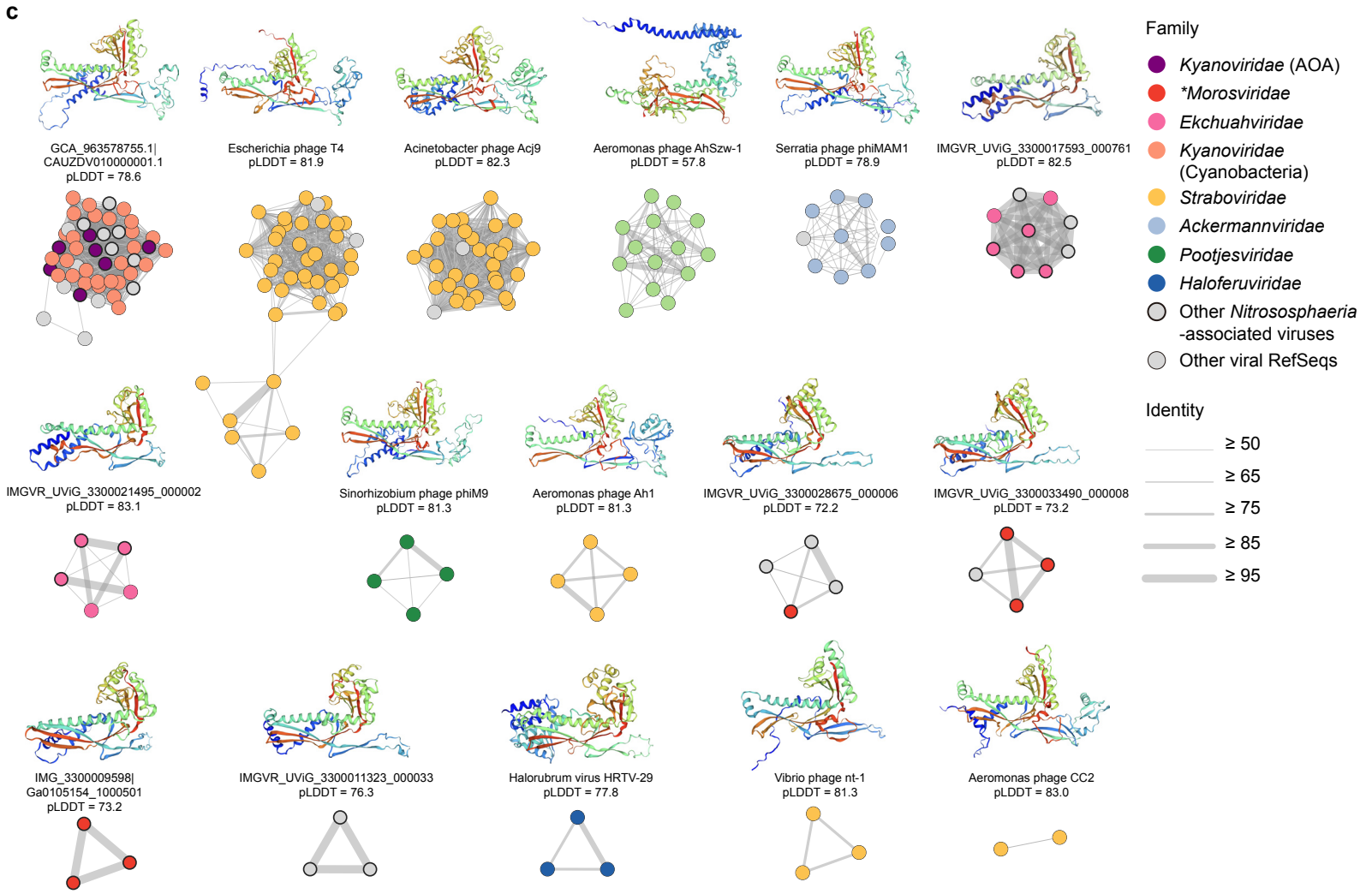

a
DUF732
PF12193
Rudiviridae
Thermoprotei
MGE_F019
MLB:79
MLB:59
MLB:100
Methanomicrobia
Ahmunviridae
MLB:73
MLB:96
*Demeterviridae
Two subunits of SIRV-like MCPs
BI convergence indicies:
minESS: 631
PSRF: 1
ASDSF: 0.01
BI vs. ML indicies:
normRF to ML: 0.15
consensus clades to ML: 49/52 (94%)
Tree scale: 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
Demerecviridae
MLB:100
Actinomycetes
α-/ß-/γ-proteobacteria
Bacilli/Actinomycetes/
α-/ß-/γ-proteobacteria
MLB:74
Bacilli/Clostridia
MLB:100
*Morosviridae
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
Flavobacteria/Rhodothermia/
Bacilli/α-proteobacteria
Klaczkowskaviridae
Malyovolviridae
MLB:98
Poolesviridae
α-proteobacteria
ε-proteobacteria
MLB:95
Straboviridae
γ-proteobacteria
γ-proteobacteria
Straboviridae
MLB:71
Ackermannviridae
MLB:49
MLB:96
Kyanoviridae
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
*Aetherviridae
Halobacteria
MLB:52
MLB:100
*Thanatosviridae
MLB:50
Bacilli/Actinomycetes/
γ-proteobacteria
MLB:100
*Morosviridae
MLB:87
α-/γ-proteobacteria
MLB:100
MLB:99
MLB:100
MLB:100
Kyanoviridae
*Erebusviridae
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
*Morosviridae
Peduoviridae
α-/γ-proteobacteria
Arenbergviridae
Deinococci
Flavobacteria/Rhodothermia
α-/γ-proteobacteria
Flavobacteria/
α-proteobacteria
Straboviridae
α-/γ-proteobacteria
Ackermannviridae
ß-/γ-proteobacteria
Kyanoviridae
α-proteobacteria/
Cyanophyceae
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

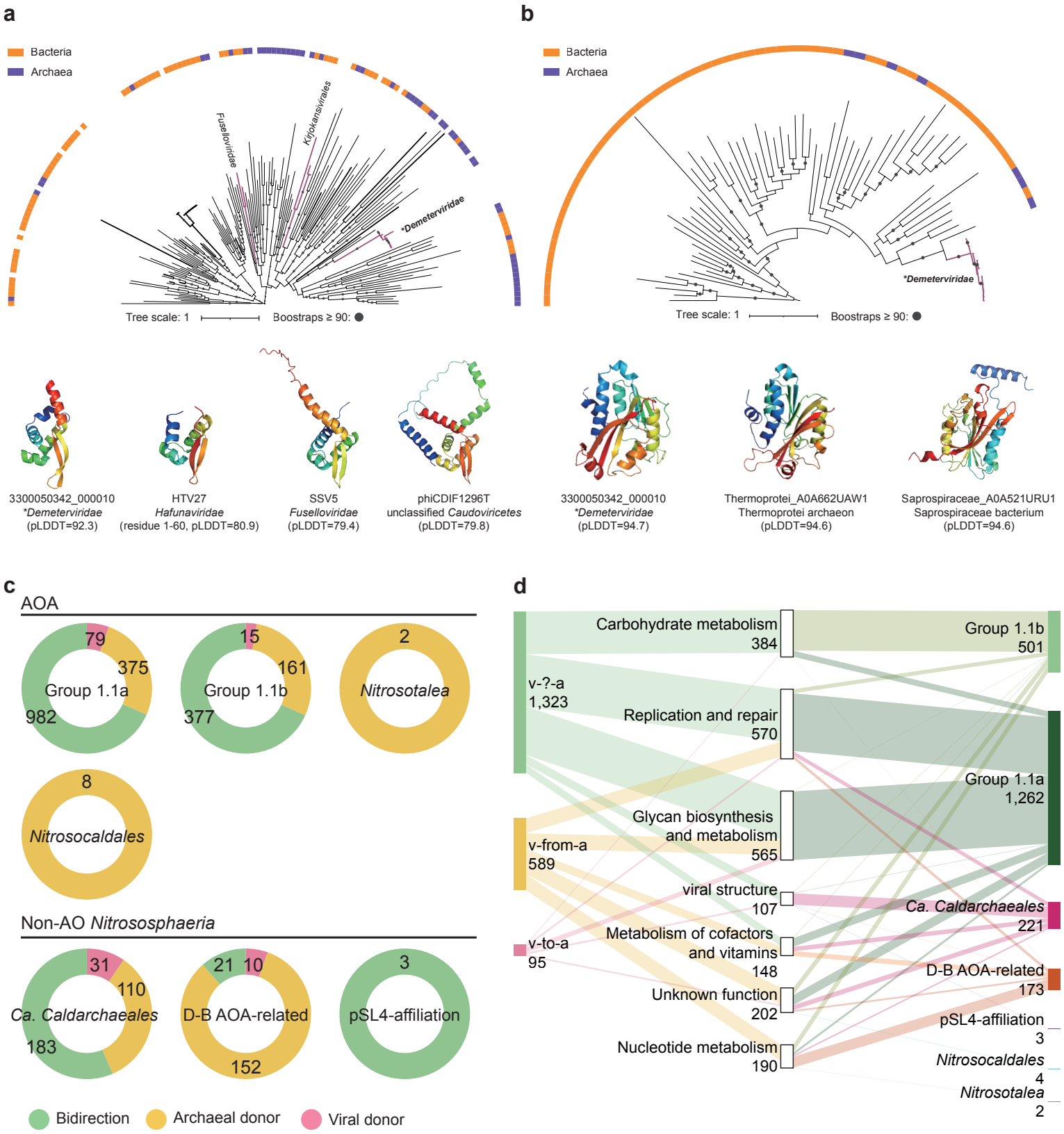


(1) IMGVR_UViG_3300025089_000008; (2) Sulfolobus islandicus rod-shaped virus 2; (3) Acidianus rod-shaped virus 2; (4) IMGVR_UViG_3300014911_000547; (5) IMG_3300025089|Ga0209394_1003436



(1) IMG_3300025089|Ga0209394_1001498; (3) Sulfolobus islandicus rod-shaped virus 2 ; (4) Saccharolobus solfataricus rod-shaped virus 1; (5) Methanophagales virus PBV300; (6) IMG_3300025089|Ga0209394_1001595




a
Bacteria
Archaea
Fuselloviridae
Kijiokansivirales
*Demeterviridae
Tree scale: 1
Boostraps ≥ 90:
b
Bacteria
Archaea
*Demeterviridae
Tree scale: 1
Boostraps ≥ 90:
3300050342_000010
*Demeterviridae
(pLDDT=92.3)
HTV27
Hafunaviridae
(residue 1-60, pLDDT=80.9)
SSV5
Fuselloviridae
(pLDDT=79.4)
phiCDIF1296T
unclassified Caudoviricetes
(pLDDT=79.8)
3300050342_000010
*Demeterviridae
(pLDDT=94.7)
Thermoprotei_A0A662UAW1
Thermoprotei archaeon
(pLDDT=94.6)
Saprospiraceae_A0A521URU1
Saprospiraceae bacterium
(pLDDT=94.6)
c
AOA
79
375
Group 1.1a
982
15
161
Group 1.1b
377
2
Nitrosotalea
8
Nitrosocaldales
Non-AO Nitrososphaeria
31
110
Ca. Caldarchaeales
183
21
10
D-B AOA-related
152
3
pSL4-affiliation
Bidirection
Archaeal donor
Viral donor
d
Carbohydrate metabolism
384
v-?-a
1,323
Replication and repair
570
Glycan biosynthesis
and metabolism
565
v-from-a
589
viral structure
107
Metabolism of cofactors
and vitamins
148
v-to-a
95
Unknown function
202
Nucleotide metabolism
190
Group 1.1b
501
Group 1.1a
1,262
Ca. Caldarchaeales
221
D-B AOA-related
173
pSL4-affiliation
3
Nitrosocaldales
4
Nitrosotalea
2