# Diversity of Bathyarchaeia viruses in metagenomes and virus-encoded CRISPR system components

Changhai Duan, Yang Liu, Ying Liu, Lirui Liu, Mingwei Cai, Rui Zhang, Qinglu Zeng, Eugene V Koonin, Mart Krupovic, Meng Li

ISME Communications, 2024, 4(1), 1–12
https://doi.org/10.1093/ismeco/ycad011
Advance access publication: 10 January 2024
Original Article

# Diversity of *Bathyarchaeia* viruses in metagenomes and virus-encoded CRISPR system components

Changhai Duan[1,2,3,4], Yang Liu[2,3], Ying Liu[5], Lirui Liu[2,3], Mingwei Cai[2,3], Rui Zhang[2,3], Qinglu Zeng[4], Eugene V. Koonin[6], Mart Krupovic[5], Meng Li[1,2,3,*]

[1]SZU-HKUST Joint PhD Program in Marine Environmental Science, Shenzhen University, Shenzhen 518060, China
[2]Archaeal Biology Center, Institute for Advanced Study, Shenzhen University, Shenzhen 518060, China
[3]Shenzhen Key Laboratory of Marine Microbiome Engineering, Institute for Advanced Study, Shenzhen University, Shenzhen 518060, China
[4]Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong 999077, China
[5]Institut Pasteur, Université Paris Cité, Archaeal Virology Unit, Paris 75015, France
[6]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

*Corresponding author: Meng Li. Email: limeng848@szu.edu.cn

## Abstract

*Bathyarchaeia* represent a class of archaea common and abundant in sedimentary ecosystems. Here we report 56 metagenome-assembled genomes of *Bathyarchaeia* viruses identified in metagenomes from different environments. Gene sharing network and phylogenomic analyses led to the proposal of four virus families, including viruses of the realms *Duplodnaviria* and *Adnaviria*, and archaea-specific spindle-shaped viruses. Genomic analyses uncovered diverse CRISPR elements in these viruses. Viruses of the proposed family "*Fuxiviridae*" harbor an atypical Type IV-B CRISPR-Cas system and a Cas4 protein that might interfere with host immunity. Viruses of the family "*Chiyouviridae*" encode a Cas2-like endonuclease and two mini-CRISPR arrays, one with a repeat identical to that in the host CRISPR array, potentially allowing the virus to recruit the host CRISPR adaptation machinery to acquire spacers that could contribute to competition with other mobile genetic elements or to inhibit host defenses. These findings present an outline of the *Bathyarchaeia* virome and offer a glimpse into their counter-defense mechanisms.

**Keywords:** Bathyarchaeia, virus, Type IV-B CRISPR-Cas system, Cas4 protein

## Introduction

*Bathyarchaeia*, formerly the Miscellaneous Crenarchaeotal Group (MCG), is an archaeal class widespread in marine and freshwater sediments [1-6]. The estimated global abundance of *Bathyarchaeia* reaches up to $2.0–3.9 \times 10^{28}$ cells, representing one of the most abundant groups of archaea on Earth [7]. Genomic analyses suggest that *Bathyarchaeia* lead an acetyl-CoA-centered heterotrophic lifestyle with the potential for acetogenesis [7], methane metabolism [8], and sulfur reduction [9]. *Bathyarchaeia* also encompass a variety of genes encoding carbohydrate-active enzymes [10] and thus likely can unitize various carbohydrates and lignin [11]. The diverse metabolic potential of *Bathyarchaeia* contributes to their predominance in sedimentary environments, rendering them essential players in the global carbon cycle [5, 6, 9, 12].

Viruses, as the most abundant biological agents on the planet [13, 14], have a major impact on the composition and activity of microbial communities [15-18]. Bacteria and archaea evolved enormously versatile repertoires of antivirus immune systems. In particular, nearly all archaea and many bacteria encode CRISPR-Cas, the prokaryotic adaptive immunity systems [19]. The CRISPR-Cas system selectively acquires foreign DNA fragments (protospacers) and stores them as spacers in the CRISPR array, which is expressed to produce CRISPR (cr) RNAs that serve as guides recognizing the DNA or RNA target and recruiting CRISPR effector nucleases [20]. The CRISPR-Cas systems provide the most reliable basis for viral host prediction by linking host spacers to the cognate virus protospacers [21].

In response to the host defenses, viruses infecting bacteria and archaea evolved a broad repertoire of counter-defense mechanisms to evade immunity [22-24], engaging in the perennial arms race. In particular, many viruses encode diverse anti-CRISPR proteins (Acrs) that specifically target different CRISPR-Cas subtypes [25, 26]. These Acrs function by directly interacting with various components of the CRISPR-Cas system [27-29] or modulating the levels of cyclic oligoadenylate [30], thereby suppressing the functionality of the CRISPR-Cas system.

To date, only a limited number of archaeal viruses have been isolated by traditional cultivation-dependent methods [31-34]. In recent years, an increasing diversity of archaeal viruses has been uncovered through metagenomic data mining [35], including viruses associated with Asgard archaea [36-38], methanogenic archaea [39], methanotrophic ANME-1 archaea [40], ammonia-oxidizing archaea [41-44], and marine Group II *Euryarchaeota* (Poseidoniales) [45]. Although viruses linked to *Bathyarchaeia* have been reported [46-49], their gene content, diversity, classification, and virus–host interactions have not comprehensively assessed,

**Received** 3 December 2023. **Revised:** 14 December 2023. **Accepted:** 18 December 2023

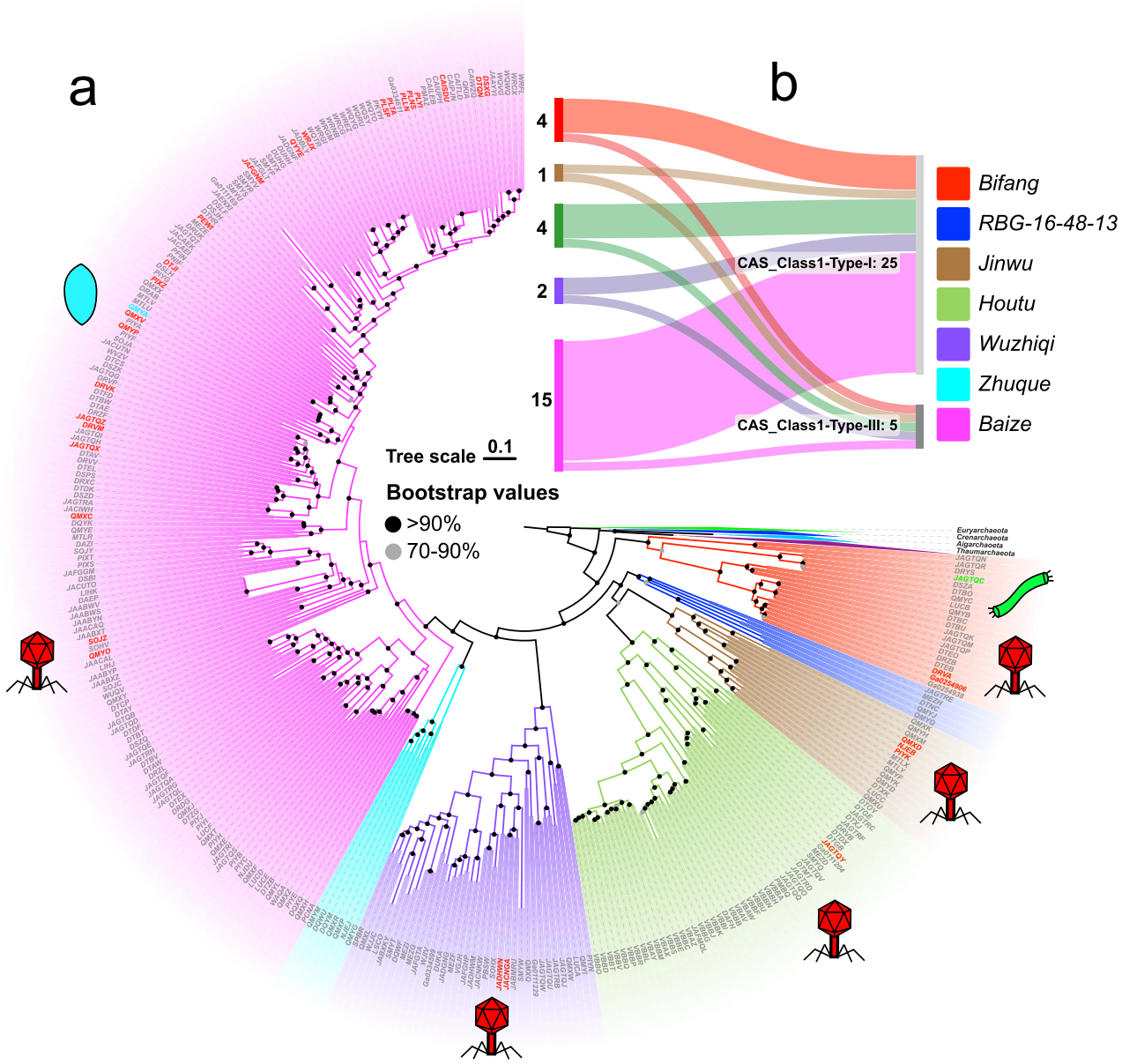


**Figure 1.** Distribution of identified viruses across the evolutionary tree of the *Bathyarchaeia*; (A) maximum likelihood tree of *Bathyarchaeia* was reconstructed based on the modified set of 51 marker genes; the outgroup was taken from Ren *et al.* [116] and includes representatives of the phyla *Euryarchaeota*, *Crenarchaeota*, *Aigarchaeota*, and *Thaumarchaeota*; *Bathyarchaeia* orders are highlighted in different background colors; viruses identified in this study are indicated with colored symbols denoting the respective virion architectures, and their corresponding hosts are highlighted with the same color; red icosahedron represents viruses in the realm *Duplodnaviria*, green rod represents viruses in the realm *Adnaviria*, and blue spindle represents the spindle-shaped virus; (B) subtypes of CRISPR-Cas systems distributed in different *Bathyarchaeia* orders: Bifang, Bifangarchaeales; Jinwu, Jinwuousiales; Houtu, Houtuarculales; Wuzhiqi, Wuzhiqiibiales; Zhuque, Zhuquarculales; Baize, Baizomonadales; only high-quality MAGs were selected.

warranting further investigation. Here, we describe the results of metagenomic analysis revealing a substantial diversity of viruses associated with *Bathyarchaeia*, some of which encode CRISPR-Cas system components and predicted Acr could potentially interact with host immunity systems.

## Results and Discussion

### Discovery of viruses associated with *Bathyarchaeia*

In total, 367 metagenome-assembled genomes (MAGs) of *Bathyarchaeia*, including 78 high-quality MAGs (Extended Data Supplementary Table S1), were obtained from our previous results [9, 50] and public databases. Phylogenetic analysis based on a concatenated alignment of an optimized set of 51 marker proteins (see Methods) yielded 7 major clades of *Bathyarchaeia* that mostly correspond to the orders in the latest taxonomy [51] (Fig. 1A and Extended Data Supplementary Fig. S1. The biome distribution survey showed that these *Bathyarchaeia* MAGs covered diverse ecosystems, including hot springs, hydrothermal vents, and mangroves, consistent with a previous 16S rRNA gene-based ecological distribution survey [5]. Moreover, some MAGs of *Bathyarchaeia* were found in freshwater sediments, soil, and the termite guts (Fig. 2). Different *Bathyarchaeia* orders also exhibit environmental preferences. For instance, Bifangarchaeales primarily occur in hot environments, whereas Zhuquarculales are only found in

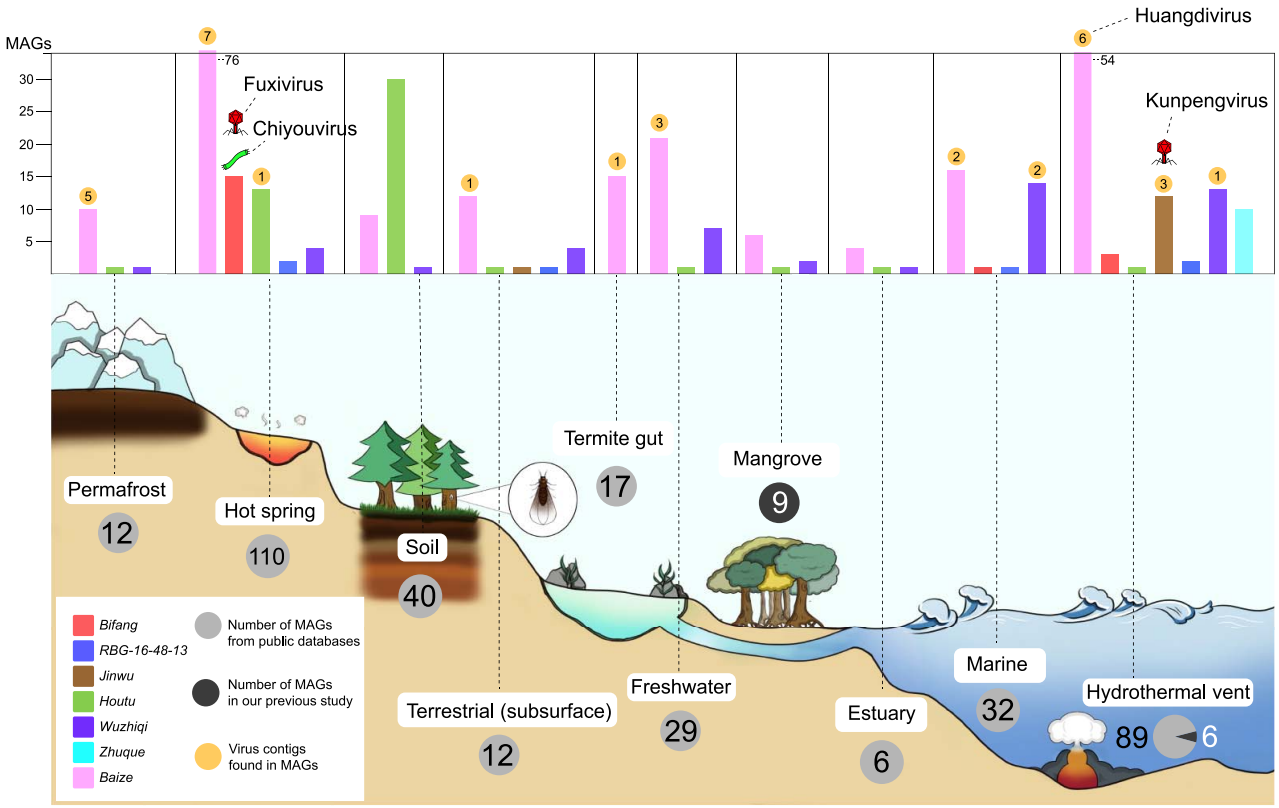


**Figure 2.** Ecological distribution of *Bathyarchaeia* and their viruses; the number of *Bathyarchaeia* orders in different types of habitats are shown by the bar chart, with the number of detected viral sequences discovered in the genomes shown in the circles above the corresponding bars; viruses identified in this study by matching CRISPR spacers (the distribution of these viruses aligns with their hosts' habitats) are marked with symbols that denote their respective virion architectures, placed above their corresponding hosts; MAGs from public databases and from our previous study are denoted by grey and black circles, respectively, with the quantity displayed inside the circles.

hydrothermal vents. By contrast, Baizomonadales are found in all environments, indicating their strong environmental adaptability.

We subsequently conducted a search for defense systems within *Bathyarchaeia* MAGs, focusing particularly on the well-studied CRISPR-Cas systems. In our analysis, we identified 39 distinct types of defense systems across 367 *Bathyarchaeia* MAGs. The three most widely distributed systems were DNA-modification systems, Phage defense candidate, and CRISPR-Cas systems, found, respectively, in 47.4%, 47.1%, and 22.9% of the genomes (Extended Data Table 1), suggesting these as the primary mechanisms of defense. In addition, commonly recognized defense systems such as restriction modification (found in 65 MAGs), AbiE (abortive infection) (found in 45 MAGs), as well as the recently discovered CBASS (found in 16 MAGs) and viperins (found in 24 MAGs) systems were also identified in a subset of *Bathyarchaeia* MAGs. The wide distribution of these defense systems suggests intense virus–host interactions in *Bathyarchaeia* (Extended Data Table 1). We then delved deeper into the CRISPR-Cas systems in high-quality *Bathyarchaeia* MAGs and identified Type I and Type III CRISPR-*cas* loci in over one-third of the MAGs. In most instances, a bathyarchaeal genome was found to carry a single CRISPR-Cas system of either Type I (17 high quality MAGs) or Type III (1 high quality MAGs). Notably, however, eight MAGs were found to encompass more than one type of CRISPR-Cas systems (Fig. 1B). To identify putative viruses of the *Bathyarchaeia*, a dataset of CRISPR spacers was compiled from all collected MAGs. After automatic screening followed by manual inspection (see Materials and methods), 49 high-confidence CRISPR array from different Bathyarchaeia orders were identified, containing 1602 spacers (Extended Data Supplementary Fig. S2). The spacers were used to search for protospacers in the viral sequences of IMG/VR v3 database [52]. Additionally, a search for potential proviruses in the host genomes was conducted (see Materials and methods). In total, 56 contigs were assigned to *Bathyarchaeia* viruses based on CRISPR spacer matches or unequivocal integration into the host genome. Among these, 54 contigs were found to encode major capsid proteins (MCPs), attesting to their viral nature [53]. Six viral contigs corresponded to complete genomes, as indicated by the presence of terminal repeats (Extended Data Supplementary Table S3).

## Four putative families of *Bathyarchaeia* viruses

Protein sharing network analysis identified four distinct groups of *Bathyarchaeia* viruses, including a large group that consisted of six viral clusters, roughly equivalent to genus-level groups [54] (Fig. 3). Additionally, we carried out a genome-wide sequence similarity comparison and examined the phylogenies of hallmark genes of *Bathyarchaeia* viruses, comparing them with known archaeal viruses (Extended Data Supplementary Figs S3–S5, see online supplementary material for a color version of this figure). Based on the results of these analyses, we identified three distinct types of viruses, encompassing four putative family-level groups. The families "*Fuxiviridae*" and "*Kunpengviridae*" include head-tailed viruses of the class *Caudoviricetes* in the realm *Duplodnaviria*. The family "*Chiyouviridae*" consists of filamentous viruses of the archaea-specific realm *Adnaviria* [55]. The fourth putative family, "*Huangdiviridae*," with only one representative genome, includes an archaea-specific spindle-shaped virus; the spindle-shaped viruses have not yet been classified at higher taxonomy ranks (Fig. 3).

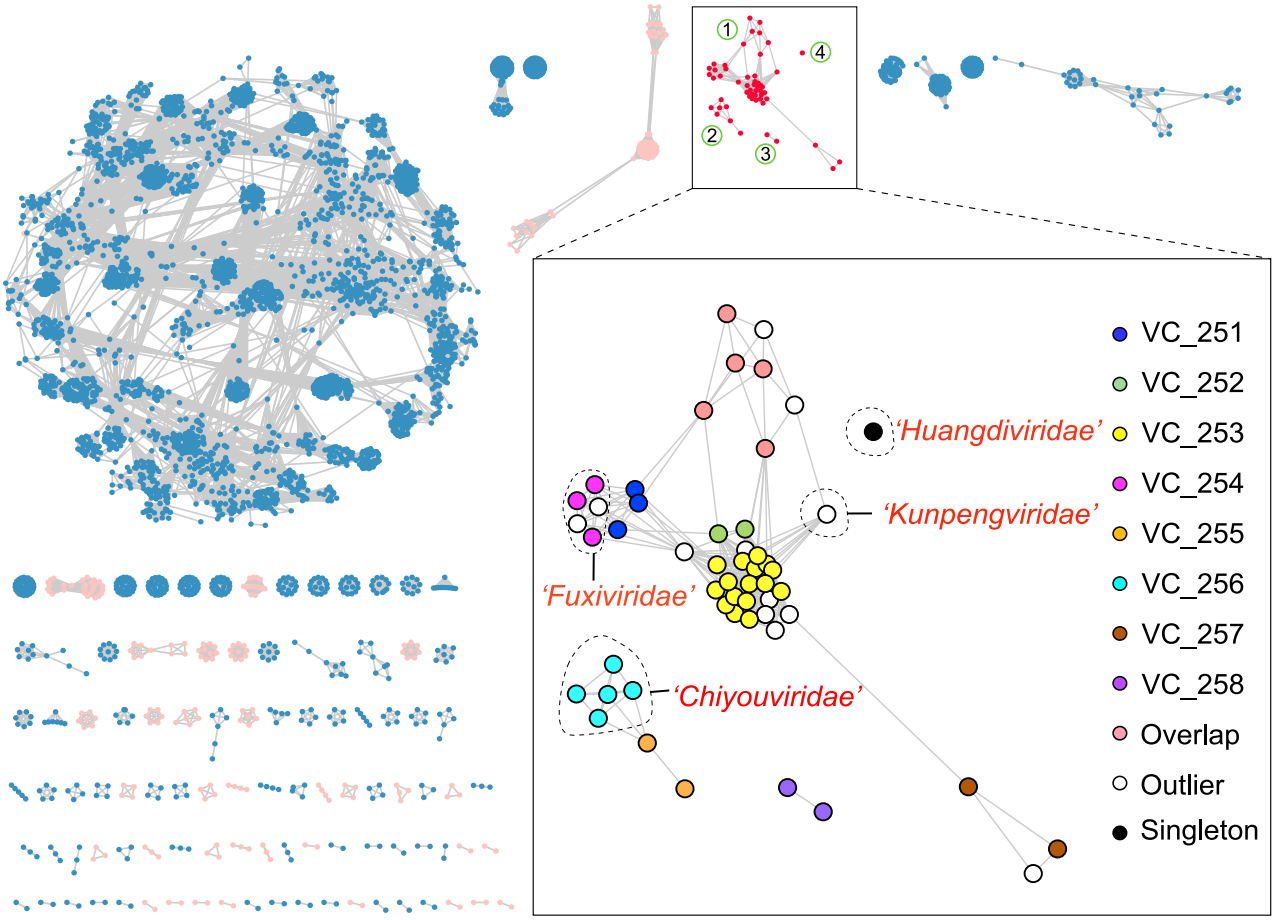


**Figure 3.** Classification of *Bathyarchaeia* viruses based on the whole-genome protein-sharing network with other prokaryotic viruses; the whole-genome protein-sharing network analysis was constructed using vConTACT2 v.0.11.3 for the taxonomic assignment of 56 *Bathyarchaeia* viral genomes; *Bathyarchaeia* virus clusters are outlined with a rectangle in the complete network; *Bathyarchaeia* viruses are assigned to four distinct groups (numbered within circles), including one large cluster; viral clusters (VCs) are indicated by the colored spheres within the inset; the proposed virus families with complete genomes are separated by dashed lines and appended with the corresponding names; the light pink and light blue clusters outside of the inset represent archaeal and bacterial viruses, respectively; the networks were visualized with Cytoscape v.3.9.1.

The proposed family "*Fuxiviridae*" is represented by three nearly identical complete genomes (Fuxivirus) from hot springs (Fig. 2) that encompass protospacers targeted by Type I-A CRISPR spacers from the *Bathyarchaeia* order Bifangarchaeales (Figs 1 and 4, Extended Data Supplementary Table S2). Fuxivirus has a smaller genome compared to the typical size of archaeal viruses of the class *Caudoviricetes* (median size of 54.3 kb, n = 44), with a length of 31 982 bp (Fig. 4A). Fuxivirus encodes all the hallmark proteins of *Caudoviricetes*, namely, a HK97-like MCP (gene Fuxivirus_34 and Extended Data Supplementary Fig. S3B), a portal protein (gene Fuxivirus_28), a terminase large subunit (LSU) (gene Fuxivirus_27), a tail tube protein (gene Fuxivirus_38), and several other tail components (Fig. 4A and Extended Data Supplementary Table S4), which are similar to those of the previously characterized archaeal tailed viruses [56]. In addition to the viral hallmark genes, Fuxivirus encodes several putative DNA-binding proteins, such as predicted transcription factors containing Zn-finger and winged-helix-like domains (Fig. 4A and Extended Data Supplementary Table S4) that likely regulate viral or host gene expression [57]. Notably, Fuxivirus also encodes CRISPR-Cas system components, namely, a Type IV *cas* gene cluster (genes Fuxivirus_1 to Fuxivirus_5) with a mini CRISPR array, and a Cas4-like protein (gene Fuxivirus_12) (Fig. 4A).

The proposed family "*Kunpengviridae*" was detected in hydrothermal vents (Fig. 2) and includes one complete viral genome (Kunpengvirus), that is targeted by a single spacer (100% match) from *Bathyarchaeia* sp. QMXD of the order Jinwuousiales (Figs 1 and 4B, Extended Data Supplementary Table S2). Kunpengvirus encodes the hallmark capsid morphogenesis proteins of *Caudoviricetes*, as well as a suite of tail proteins including the baseplate protein (Fig. 4B and Extended Data Supplementary Table S4). Kunpengvirus also encodes an integrase (gene Kenpengvirus_28), suggesting that it can integrate into the host genome as a provirus. In addition, this virus encompasses genes for a deoxynucleoside monophosphate kinase (gene Kenpengvirus_41), an MCM-like helicase (gene Kenpengvirus_46) and a family B DNA polymerase (gene Kenpengvirus_42) (Fig. 4B and Extended Data Supplementary Table S4), indicative of (at least, partially) autonomous genome replication [58]. Notably, unlike in other tailed viruses, the proofreading DEDDy 3′-5′ exonuclease (gene Kenpengvirus_45) and the family B DNA polymerase domains are encoded by two distinct genes. Additionally, Kunpengvirus encodes a homolog of ribosomal protein bL19 (gene Kunpengvirus_2), which is typically present in bacteria and eukaryotes (chloroplasts and mitochondria) but not in archaea, except for *Candidatus Aenigmarchaeota*. The bL19 is located at the 30S–50S ribosomal subunit interface and is thought to contribute to the structure and function of the aminoacyl-tRNA binding site [59]. A number of ribosomal proteins, including bL19, were previously identified in bacterial viruses [60], but the only ribosomal protein so far detected in

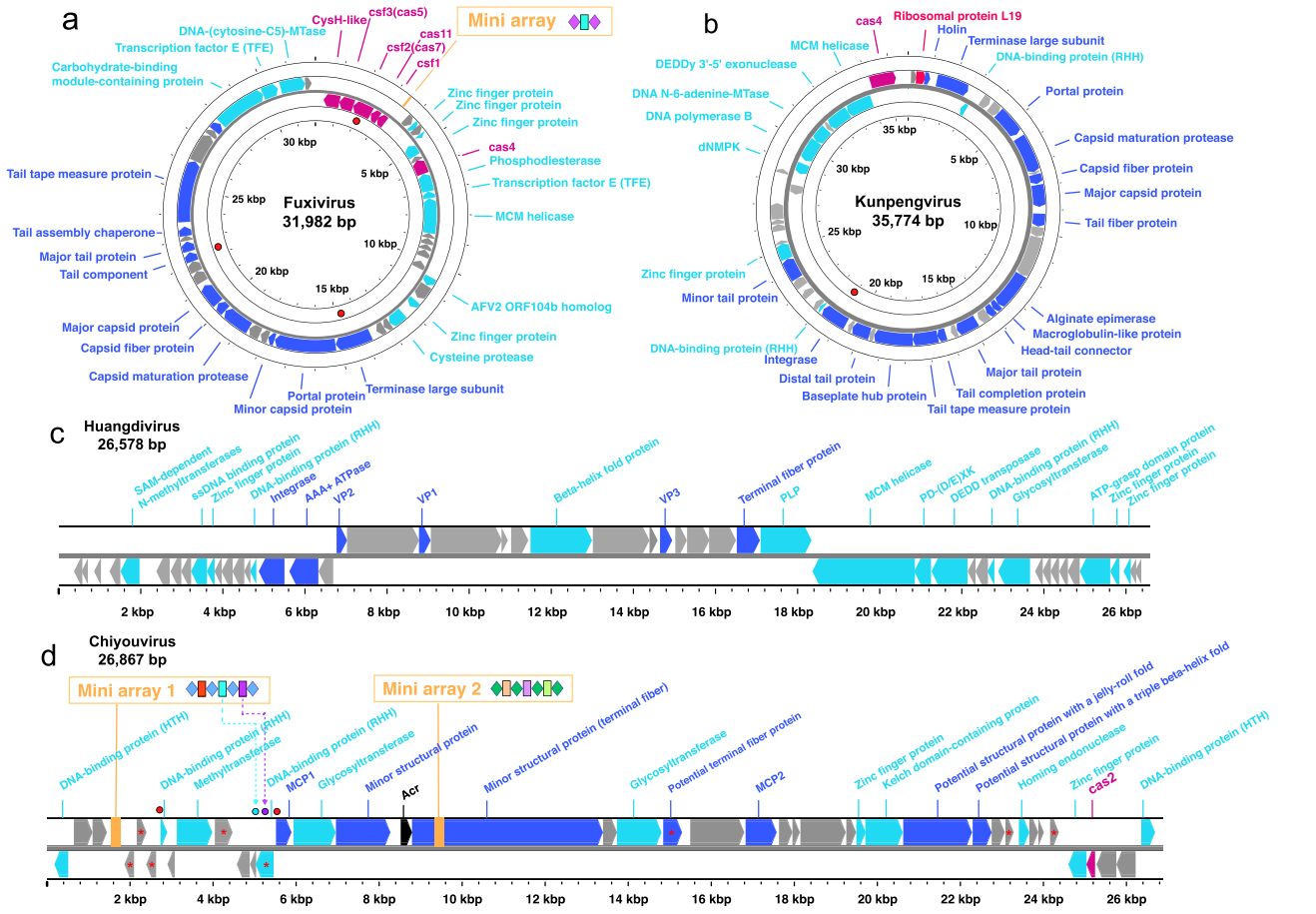


**Figure 4.** Maps of complete genomes of *Bathyarchaeia* viruses; (A) Genome map of Fuxivirus; (B) genome map of Kunpengvirus; (C) genome map of Huangdivirus; (D) genome map of Chiyouvirus; genes annotated by HHblits with a probability greater than 95% are shown as different colors; genes related to Type IV B CRISPR-Cas system, Cas4, and Cas2 are indicated in rose pink, mini-CRISPR array in vibrant orange, genes specific to viruses in dark blue, predicted Acr in black, other annotated genes in light blue; the positions of targeted protospacers are indicated with red circle; the organization of CRISPR mini-arrays is shown above the genome maps; in Chiyouvirus CRISPR Array 1, the self-targeting spacers are highlighted in light blue and purple; their corresponding target sites on the genome are marked with circles in the same color; transmembrane proteins of Chiyouvirus predicted using CCTOP [117] server are indicated with red asterisks; fdetailed gene annotations are in Supplementary Table S4.

archaeal viruses had been L21e [56]. The presence of bL19 in Kunpengvirus suggests that modification of the host translation apparatus by archaeal viruses is more common than currently recognized. Similar to the Fuxivirus, Kunpengvirus also encodes a Cas4-like endonuclease (Fig. 4B and gene Kenpengvirus_47).

A unique viral genome (Huangdivirus) representing the proposed family "*Huangdiviridae*" was identified as an apparent provirus in *Bathyarchaeia* sp. QMYA of the order Baizomonadales from a deep-sea hydrothermal vent (Figs 1 and 2). Sensitive sequence comparison using HHblits identified three virus-encoded structural proteins (VPs) VP1–3 (Fig. 4C and Extended Data Supplementary Table S4), homologous to the structural proteins of archaeal spindle-shaped viruses of the family *Fuselloviridae*, as well as an AAA+ ATPase (gene Huangdivirus_17, HHblits best hit to ATV ATPase, with 99.8% probability). Structural predictions for VP1 (gene Huangdivirus_21) and VP3 (gene Huangdivirus_28) indicated that both proteins contain two hydrophobic $\alpha$-helices connected by a short turn (Extended Data Supplementary Fig. S4A) resembling the typical structure of the MCPs of spindle-shaped viruses [61]. Huangdivirus VP2 (gene Huangdivirus_19) is most closely related to the viral DNA-binding protein VP2 of Sulfolobus spindle-shaped virus 1 (SSV1) [62] (HHblits probability of 99.49%). As in the case of SSV1 [62], we detected the consensus glycosylation motifs (N-X-S/T) in Huangdivirus VP1 and VP3, which may be glycosylated by the virus-encoded glycosyltransferase (gene Huangdivirus_40 and Extended Data Supplementary Fig. S5B). Capsid protein glycosylation is thought to contribute to the virion stability, particularly for viruses infecting hyperthermophilic hosts [62]. Similar to fuselloviruses infecting hyperthermophilic archaea of the order Sulfolobales [63], Huangdivirus encodes a putative integrase of the tyrosine recombinase superfamily (gene Huangdivirus_16; Fig. 4C), which is likely to be responsible for viral DNA integration into the host chromosome. Notably, Huangdivirus encodes a predicted polysaccharide lyase (Fig. 3C and gene Huangdivirus_33) that might be involved in the break-down of the S-layer polysaccharides [64, 65] of host cell wall during infection.

"*Chiyouviridae*" is a potential new family of filamentous viruses in the order *Ligamenvirales* (class *Tokiviricetes*, realm *Adnaviria*). "*Chiyouviridae*" was detected in a hot spring and represented by one complete viral genome (Chiyouvirus; Fig. 2) targeted by two spacers of *Bathyarchaeia* sp. JAGTQC in the order Bifangarchaeales (Figs 1 and 4D). Phylogenomic analysis of all available *Tokiviricetes* [66] genomes recapitulated the previously established relationships and showed that Chiyouvirus forms

a separate clade within the order *Ligamenvirales*, most closely related to the families *Rudiviridae* and *Ungulaviridae* (Extended Data Supplementary Fig. S5A). Whole proteome comparison showed <50% average amino acid identity (AAI) between protein homologs from Chiyouvirus and members of other viral families, with the highest AAI (46%) with the genus *Icerudivirus* of the family *Rudiviridae* (Extended Data Supplementary Fig. S5B). Similar to all members of the families *Ungulaviridae* and *Lipothrixviridae* but only some members of the family *Rudiviridae* [66], Chiyouvirus encodes two MCPs (genes Chiyouvirus_14 and Chiyouvirus_23), each comprising an alpha-helix bundle (Fig. 4D and Extended Data Supplementary Fig. S5C). In rudiviruses, unlike other members of the realm *Adnaviria*, the second MCP paralog, when present, is not incorporated into virions [67]. Thus, it remains unclear whether both Chiyouvirus MCPs are involved in virion formation. In addition, Chiyouvirus encodes a large minor structural protein (Fig. 4D and gene Chiyouvirus_18), a homolog of SIRV2 P1070, which is thought to be involved in the formation of the virion terminal filaments that are responsible for host recognition [68, 69]. Predictions for transmembrane proteins indicated that Chiyouvirus encodes eight potential transmembrane proteins (Fig. 4D) suggesting that it is a membrane-enveloped filamentous virus, similar to the adnaviruses in the families *Lipothrixviridae*, *Ungulaviridae*, and *Tristromaviridae* [70].

## CRISPR-Cas systems and potential counter defense mechanisms in *Bathyarchaeia* viruses

We explored the Type IV CRISPR-Cas system encoded by "*Fuxiviridae*" in greater detail, including phylogenetic analysis and gene locus comparison. The viral CRISPR-*cas* locus encodes all typical components of the Type IV CRISPR-Cas effector module (Fig. 4A and Extended Data Supplementary Table S4), in particular, the signature protein Csf1 (gene Fuxivirus_5), the apparent LSU of the effector complex, but no adaptation module. Phylogenetic analysis of Cas proteins, including Csf2 (Cas7) (gene Fuxivirus_3), the most conserved protein in the Type IV systems [71], showed that CRISPR-Cas system of "*Fuxiviridae*" belongs to the IV-B subtype (Extended Data Supplementary Fig. S6). Additionally, a CysH-like protein (gene Fuxivirus_1), which is tightly associated with the Type IV-B systems, is encoded in the virus genome adjacent to Csf3 (Cas5) (gene Fuxivirus_2; Fig. 4A). Phylogenetic analysis showed that Fuxivirus CysH-like protein does not belong to the viral CysH branch but is rather associated with CysH-like proteins from other Type IV-B systems (Extended Data Supplementary Fig. S6C). Previously, Type IV CRISPR-Cas systems have been observed in plasmids and prophages [72, 73] as well as several lytic phages [74], but not in archaeal viruses.

Most Type IV-B systems lack both the adaptation module and a CRISPR array, but the Type IV-B CRISPR-*cas* locus in the Fuxivirus genome contains a CRISPR mini-array that consists of two repeats and a spacer (Fig. 4A). A Type IV-B system has been reported to form a filamentous RNP complex, which predominantly assembles on non-CRISPR RNAs, without apparent sequence specificity, suggesting a function distinct from adaptive immunity [75]. However, given the presence of the mini-array, we hypothesize that the Fuxivirus Type IV-B complex could incorporate spacers, conceivably, by recruiting the host adaptation machinery, and utilize a virus-encoded crRNA to target host DNA or other coinfecting mobile genetic elements (MGEs), thus, being potentially involved in evading host immunity and/or inter-MGE conflicts [76]. However, no full-length protospacer matches for the Fuxivirus mini-array spacer sequence were found in either the host or other MGEs, so that further validation of the counter-defense function of the Fuxivirus Type IV-B CRISPR-Cas system is needed.

Although we could not detect a complete CRISPR-Cas system in the partial genome of the inferred Fuxivirus host, *Bathyarchaeia* sp. DRVA contains two CRISPR arrays and two *cas3* gene that are not adjacent to the CRISPR arrays. Notably, a close relative of *Bathyarchaeia* sp. DRVA, *Bathyarchaeia* sp. JAGTQM, from the same order Bifangarchaeales, contains complete Type I-A and III-D CRISPR-Cas systems and was found to share identical CRISPR repeat and a closely similar Cas3 protein (84.03% identity, 100% coverage) with *Bathyarchaeia* sp. DRVA. These observations suggest the presence of a CRISPR-Cas system(s) in the DRVA genome as well. Notably, we found that the *cas7* gene of Fuxivirus was targeted by a host spacer (Fig. 4A), suggesting that the virus-encoded Type IV-B system, and with it, possibly, the virus reproduction, can be inhibited by the host CRISPR-Cas systems.

The Cas4 homologs encoded by "*Fuxiviridae*" (gene Fuxivirus_12) and "*Kunpengviridae*" (gene Kenpengvirus_47) could represent an additional counter-defense factor. Cas4 is a P-D/ExK family nuclease that is a common component of CRISPR-Cas systems that assists Cas1–Cas2 integration complexes in the acquisition of CRISPR spacers [77], but many Cas4 homologs are encoded outside CRISPR-*cas* loci [78]. Phylogenetic analysis showed that these *Bathyarchaeia* viral Cas4-like proteins are most closely related to Cas4 homologs encoded by Sulfolobus-infecting rudiviruses (Extended Data Supplementary Fig. S6B), suggesting the possibility of horizontal gene transfer between unrelated archaeal viruses. Multiple sequence alignment confirmed that two Cas4 homologs encoded by "*Fuxiviridae*" and "*Kunpengviridae*" share nearly all conserved amino acids with the rudivirus SIRV2-encoded Cas4 homolog and are thus predicted to be active nucleases (Extended Data Supplementary Fig. S7). Previous studies have demonstrated that overexpression of the SIRV2-encoded Cas4 in the archaeal host substantially reduced the efficiency of exogenous spacers acquisition by the host CRISPR-Cas system [79]. Thus, the Fuxivirus and Kunpengvirus Cas4 might inhibit spacer acquisition by the CRISPR systems of Bathyarchaeia hosts. However, involvement of this nuclease in the virus genome replication cannot be ruled out either [80].

We identified two mini-arrays in Chiyouvirus genome, each containing four repeats and three spacers (Fig. 4D), and the repeats in Chiyouvirus mini-array 1 are identical to the repeats in the host CRISPR array. Thus, the pre-crRNA transcribed from this viral mini-array is predicted to be processed by the host CRISPR-Cas system and the mature viral crRNA would remain bound by the host effector complex [72]. Furthermore, the virus could potentially recruit the host CRISPR adaptation machinery to incorporate additional spacers into the mini-array and employ the respective crRNAs in inter-MGE competition, as demonstrated for some bacterial and archaeal micro-array carrying viruses [72, 76], or for abrogation of host defenses. However, intriguingly, in the Chiyouvirus mini-array 1, we identified two self-targeting spacers, both with a 100% match to the corresponding protospacers (Fig. 4D). The role of these self-targeting spacers remains unclear. For the spacers in the Chiyouvirus mini-array 2, no potential targets were identified.

In addition to the mini-arrays, Chiyouvirus encodes a homolog of Cas2 nuclease (gene Chiyouvirus_39), with a conserved $Mg^{2+}$ binding site (Extended Data Supplementary Fig. S8), which is an essential structural subunit of the adaptation complex in CRISPR-Cas systems. The virus-encoded Cas2 homolog might be a dominant negative inhibitor of spacer acquisition by the host CRISPR-Cas system.

Additionally, we attempted to predict Acrs among the *Bathyarchaeia* viral proteins by using a recently developed deep learning method [81]. We found that the structural model of a predicted Chiyouvirus Acr protein (gene Chiyouvirus_17) was significantly similar to the N-terminal domain of AcrIF24 (7DTR, chain A) (Extended Data Supplementary Fig. S9), which inhibits the activity of a Type I-F CRISPR-Cas system by forming a dimer and inducing the dimerization of the Csy complex, blocking the hybridization of target DNA to crRNA [82].

Finally, in addition to the components of CRISPR-Cas systems, Fuxivirus and Kunpengvirus encode stand-alone DNA methyltransferases (genes Fuxivirus_47 and Kenpengvirus_43) that could provide protection against the host restriction-modification systems.

To conclude, in this work, we identify four distinct groups of *Bathyarchaeia* viruses that can be expected to become new viral families. Notably, these viruses encode various components of CRISPR-Cas systems that could interfere with the host CRISPR-Cas immunity and/or mediate inter-virus conflicts. Thus, this study provides a glimpse into the virome of a widespread, comparatively abundant but poorly characterized class of archaea and the virus–host interactions in these organisms.

## Material and methods

### *Bathyarchaeia* genome collection and classification

*Bathyarchaeia* genomes were obtained from our previously reported dataset [9, 50] and the NCBI (https://www.ncbi.nlm.nih.gov) and IMG/M (https://jgi.doe.gov) public database with key words "*Bathyarchaeota*," "*Bathyarchaeia*," and "MCG" (miscellaneous *Crenarchaeota* group) (up until 30 June 2021). *Bathyarchaeia* genome classification was based on 51 high-coverage marker gene optimized from GTDB [83] v207 (TIGR01171 and TIGR02389 were excluded due to the low coverage). Marker genes were initially identified with GTDB-Tk v2.1.1 [84] using "identify" method. Only MAGs contain >50% of the marker genes were kept. Sequences were then aligned using MAFFT-LINSI v7.457 [85] and trimmed with TrimAl v1.4 [86] (–gappyout). The maximum likelihood phylogenomic tree for concatenated 51 proteins was constructed with IQ-TREE 2.0.6 [87] (best-fit model LG + F + R13, -B 1000, –alrt 1000). The quality, contamination, GC content, and other sequence information of *Bathyarchaeia* MAGs were assessed by CheckM v1.2.2 (default setting) [88]. For each *Bathyarchaeia* MAG, the 23S, 16S, and 5S rRNA genes were initially identified and extracted using Barrnap 0.9 (https://github.com/tseemann/barrnap). Then followed by a manual curation step involving BLASTN checks against the SILVA 138.1 LSU Reference database and the SILVA 138.1 Small Subunit Reference database (https://arbsilva.de/). The tRNA identification was performed using tRNAscan-SE [89]. Biome information was extracted from the GenBank file or relevant literatures.

To delineate *Bathyarchaeia* orders, the maximum likelihood tree based on 51 proteins was converted to an ultra-metric tree and the relative evolutionary divergence (RED) [90] was calculated. An order was called when its branch length in the ultrametric tree corresponded to a RED value within the range of 0.6 ± 0.1.

### *Bathyarchaeia* defense-systems detection and CRISPR array validation

Defense systems (including CRISPR-Cas systems) in *Bathyarchaeia* MAGs were detected using PADLOC [91] (v2) and DefenseFinder [92]. Subtypes of detected CRISPR-Cas systems were classified using CRISPRCasTyper [93]. CRISPR array in *Bathyarchaeia* MAGs were searched with MinCED [94] (Default setting). To remove ambiguous CRISPR array that might not belong to *Bathyarchaeia*, all CRISPR array-containing contigs encoding proteins were searched against the nr database (2022–09) using Diamond [95] (2.1.6) BLASTP (–evalue 1e-5, —more-sensitive), and only contigs with at least three proteins assigned to *Bathyarchaeia* were retained. The validated CRISPR repeats were clustered with 90% identity using an all-against-all BLASTN [96] search (–evalue 1e-5, word size 7).

### *Bathyarchaeia* virus identification and annotation

We first carried out the CRISPR-based virus-host assignments. Spacers from validated CRISPR array were used as a nucleotide database to compare to viral contigs in the IMG/VR v3 [52] database using BLASTN with parameters (–task blastn-short -evalue 1e-5). Only viral contigs harboring protospacers with 100% coverage and a maximum of one mismatch were considered as *Bathyarchaeia* viruses.

In addition, some viruses might form proviruses that cannot be detected by CRISPR-protospacer search. Therefore, we used VirSorter2 v.2.2.3 [97] to identified potential viral sequences in *Bathyarchaeia* MAGs. Specifically, all the *Bathyarchaeia* contigs were filtered using VirSorter2 (—include-groups dsDNAphage, ssDNA —min-length 5000 —min-score 0.5) and only contigs with identifiable viral structural proteins and at least one ORF identified through a BLASTP search against the nr database as *Bathyarchaeia* were retained. Virial genome completeness estimation and the potential host regions trimming were conducted by CheckV 0.8.1 [98]. Coding sequences were identified using Prokka [99] with parameters (—kingdom viruses —gcode 1).

All the *Bathyarchaeia* viral proteins identified as describe above were annotated with DRAM [100] (viral mode) and HHblits [101] (MSA generated with UniRef30_2020_06 with three interactions, Evalue 1e-6, MSA was compared against the PDB_mmCIF70_14_Apr, SCOPe70_2.07, Pfam-A_v35 and UniProt-SwissProt-viral70_3_Nov_2021 databases with parameters -Z 250 -loc -z 1 -b 1 -B 250 -ssm 2 -sc 1 -seq 1 -dbstrlen 10 000 -norealign -maxres 32 000). All the viral genome maps were visualized using Proksee [102].

### *Bathyarchaeia* viral taxonomy assignment

To determine the taxonomic status of *Bathyarchaeia* viruses, the whole genome network analysis was carried out with vConTACT2 [54] (default parameters) against the Viral RefSeq v207 database as well as archaeal viruses proposed in the latest International Committee on Taxonomy of Viruses Report (VMR_MSL38_v1). The resulting networks were visualized using Cytoscape v3.9.1 [103] with an edge-weighted spring embedded model.

For viruses in the realm *Duplodnaviria*, proteome-scale phylogeny was constructed using the VipTree version 3.4 [104]. The analysis was carried out with archaeal virus families in the realm *Duplodnaviria*, according to the latest ICTV taxonomy classification (25 April 2023). For filamentous viruses, proteome-scale phylogeny was constructed by using VICTOR [105] with all known members of the *Tokiviricetes* class. Structural modeling of viral proteins was performed with AlphaFold2 [106] via ColabFold v1.5.1 [107] with pdb70 template mode.

### *Bathyarchaeia* viral CRISPR-Cas systems classification

All the Cas proteins were manually search with HHblits (see above). To determine the CRISPR-Cas system subtype of '*Fuxiviridae*', phylogenetic tree of all the Cas proteins was constructed with reference sequences. For each of the Cas4, Cas7, Cas5, Cas11, and

Csf1 proteins, as well as the phosphoadenosine phosphosulfate reductase (CysH) domain-containing proteins, we applied a consistent analysis procedure. Each protein type was aligned with its respective reference sequences [73, 79] (29 for Cas4, 131 for Cas7, 133 for Cas5, 78 for Cas11, 126 for Csf1, and 322 for CysH, including bona fide CysH and Cas-associated CysH-like proteins recalled from 18 988 public reference complete viral genomes [108]) using MUSCLE 5.1 [109]. The alignments were subsequently trimmed by TrimAl v1.4 (–gappyout for Cas4, Cas5, Cas11, Csf1, and CysH, –gt 0.7 for Cas7), and dropping sequences with >50% gaps. Maximum likelihood trees were then computed for each protein type using IQ-TREE 2.0.6, with the best model selected by model finder [110] (-MFP) varying depending on the proteins (LG + I + G4 for Cas4, VT + R10 for Cas7 and Cas5, VT + R3 for Cas11, WAG+I + G4 for Csf1, and WAG+F + R7 for CysH).

All final phylogenetic trees were visualized by tvBOT [111].

### *Bathyarchaeia* viral CRISPR spacer-protospacer match analysis

CRISPR array in *Bathyarchaeia* viral genome were detected with MinCED (-minNR 1). All the spacers obtained from *Bathyarchaeia* virus were BLASTN search (–task blastn-short) with viral database IMG/VR v3 and plasmid sequences database PLSDB database [112] (v. 2021_06_23_v2) as well as *Bathyarchaeia* MAGs in this study. Direct repeats were searched against CRISPR-CAS++ databases [113] and *Bathyarchaeia* CRISPR repeat dataset with BLASTN (–evalue 0.01).

### Identification of *Bathyarchaeia* viral Acrs

The potential Acrs of *Bathyarchaeia* viruses were first predicted using DeepAcr [81]. The protein structures of predicted Acrs, along with Acrs from reference [114], were modeled using AlphaFold2 by ColabFold v1.5.2 in pdb70 template mode. These potential *Bathyarchaeia* virus Acrs were then used as query models for alignment to the reference Acr models using TM-align [115], and only models with a TM score >0.3 were retained. This initial computational analysis was further supplemented with comprehensive meticulous manual inspection of the alignments and structural models.

### Etymology

"*Fuxiviridae*": Derived from Fuxi, a legendary figure in Chinese mythology known for his diverse talents and abilities. This alludes to the possibility that the virus has multifaceted counter-defense systems, capable of employing various strategies to evade the host's immune response.

"*Kunpengviridae*": Named after Kunpeng, a mythical creature in Chinese mythology known for its transformative abilities. The name alludes to the virus's ability to integrate into the host genome to form proviruses.

"*Huangdiviridae*": Named after Huangdi, the legendary Chinese sovereign often associated with important inventions. Given Huangdi's connections (the host MAG of Huangdivirus in the Baizomonadales order) in Chinese mythology.

"*Chiyouviridae*": Inspired by Chiyou, a symbol of war and invention in Chinese mythology.

## Acknowledgements

We would like to express our deepest gratitude to the following professors for their generous support and data sharing, which has been invaluable to our research. Special thanks go to Professors, Simon Roux, Brett Baker, Natasha Ivanova, Robert Kelly, Tanja Woyke, William P. Inskeep, Jim Fredrickson, Roland Hatzenpichler, Brian P. Hedlund, and Ramunas Stepanauskas for their enthusiastic assistance and valuable suggestions. We sincerely appreciate their contributions. Also, thanks to Dr Jie Pan, Mr Chengxiang Gu, and Mr Jinquan Li for their invaluable guidance and assistance.

## Author contributions

Meng Li and Changhai Duan designed the experiments; Changhai Duan, Yang Liu, and Mingwei Cai collected samples and analyzed *Bathyarchaeia* MAGs. Changhai Duan, Yang Liu, Ying Liu, Rui Zhang, Qinglu Zeng, and Mart Krupovic analyzed viral sequences. Changhai Duan, Mart Krupovic, and Eugene V. Koonin analyzed viral CRISPR-Cas systems elements. Changhai Duan, Yang Liu, Rui Zhang, Qinglu Zeng, Mart Krupovic, Eugene V. Koonin, and Meng Li wrote the paper with input from all the authors.

## Supplementary material

Supplementary material is available at *ISME Communications* online.

## Conflicts of interest

The authors declare no competing interests.

## Funding

National Key Research and Development Program of China (2022YFA0912200), the National Natural Science Foundation of China (32225003, 92251306, 31970105), the Innovation Team Project of Universities in Guangdong Province (No. 2020KCXTD023), the Shenzhen Science and Technology Program (JCYJ20200109105010363), and Shenzhen University 2035 Program for Excellent Research (2022B002) and Intramural Research Program of the National Insitutes of Health of the USA (National Library of Medicine) to E.V.K.

## Data availability

All the *Bathyarchaeia* assembled genomes ($n = 367$) used in this study were collected from publicly available databases including the NCBI (https://www.ncbi.nlm.nih.gov) and IMG/M (https://jgi.doe.gov). The accession number for the MAGs is available in Supplementary Table S1. All the viral contigs ($n = 56$) analyzed in this study were collected from the above *Bathyarchaeia* genomes and publicly available database IMG/VR v3 (https://img.jgi.doe.gov/vr). The GenBank accession number of proposed viral family and datasets generated in this study (i.e. viral contigs, protein files for the viruses, alignments, and tree files) can be accessed at [https://doi.org/10.6084/m9.figshare.24540121.v1].

## Code availability

No custom code was used.

## References

1. Inagaki F, Nunoura T, Nakagawa S *et al.* Biogeographical distribution and diversity of microbes in methane hydrate-bearing deep marine sediments on the Pacific Ocean margin. *Proc Natl Acad Sci U S A* 2006;**103**:2815–20. https://doi.org/10.1073/pnas.0511033103.

2. Kubo K, Lloyd KG, F Biddle J *et al.* Archaea of the miscellaneous Crenarchaeotal group are abundant, diverse and widespread in marine sediments. *ISME J* 2012;**6**:1949–65. https://doi.org/10.1038/ismej.2012.37.
3. Lloyd KG, Schreiber L, Petersen DG *et al.* Predominant archaea in marine sediments degrade detrital proteins. *Nature* 2013;**496**:215–8. https://doi.org/10.1038/nature12033.
4. Lazar CS, Biddle JF, Meador TB *et al.* Environmental controls on intragroup diversity of the uncultured benthic archaea of the miscellaneous Crenarchaeotal group lineage naturally enriched in anoxic sediments of the White Oak River estuary (North Carolina, USA). *Environ Microbiol* 2015;**17**:2228–38. https://doi.org/10.1111/1462-2920.12659.
5. Zhou Z, Pan J, Wang F *et al. Bathyarchaeota*: globally distributed metabolic generalists in anoxic environments. *FEMS Microbiol Rev* 2018;**42**:639–55. https://doi.org/10.1093/femsre/fuy023.
6. Fillol M, Auguet JC, Casamayor EO *et al.* Insights in the ecology and evolutionary history of the miscellaneous Crenarchaeotic group lineage. *ISME J* 2016;**10**:665–77. https://doi.org/10.1038/ismej.2015.143.
7. He Y, Li M, Perumal V *et al.* Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nat Microbiol* 2016;**1**:16035. https://doi.org/10.1038/nmicrobiol.2016.35.
8. Evans PN, Parks DH, Chadwick GL *et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. *Science* 2015;**350**:434–8. https://doi.org/10.1126/science.aac7745.
9. Pan J, Zhou Z, Béjà O *et al.* Genomic and transcriptomic evidence of light-sensing, porphyrin biosynthesis, Calvin-Benson-Bassham cycle, and urea production in Bathyarchaeota. *Microbiome* 2020;**8**:43. https://doi.org/10.1186/s40168-020-00820-1.
10. Qi YL, Evans PN, Li YX *et al.* Comparative genomics reveals thermal adaptation and a high metabolic diversity in "*Candidatus Bathyarchaeia*". *mSystems* 2021;**6**:e00252–21.
11. Yu T, Wu W, Liang W *et al.* Growth of sedimentary Bathyarchaeota on lignin as an energy source. *Proc Natl Acad Sci U S A* 2018;**115**:6022–7. https://doi.org/10.1073/pnas.1718854115.
12. Zhou Z, Liu Y, Lloyd KG *et al.* Genomic and transcriptomic insights into the ecology and metabolism of benthic archaeal cosmopolitan, Thermoprofundales (MBG-D archaea). *ISME J* 2019;**13**:885–901. https://doi.org/10.1038/s41396-018-0321-8.
13. Chevallereau A, Pons BJ, van Houte S *et al.* Interactions between bacterial and phage communities in natural environments. *Nat Rev Microbiol* 2022;**20**:49–62. https://doi.org/10.1038/s41579-021-00602-y.
14. López-García P, Gutiérrez-Preciado A, Krupovic M *et al.* Metagenome-derived virus-microbe ratios across ecosystems. *ISME J* 2023;**17**:1552–63. https://doi.org/10.1038/s41396-023-01431-y.
15. Jiao N, Herndl GJ, Hansell DA *et al.* Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nat Rev Microbiol* 2010;**8**:593–9. https://doi.org/10.1038/nrmicro2386.
16. Brussaard CPD, Wilhelm SW, Thingstad F *et al.* Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J* 2008;**2**:575–8. https://doi.org/10.1038/ismej.2008.31.
17. Fang X, Liu Y, Zhao Y *et al.* Transcriptomic responses of the marine cyanobacterium *Prochlorococcus* to viral lysis products. *Environ Microbiol* 2019;**21**:2015–28. https://doi.org/10.1111/1462-2920.14513.
18. Xiao X, Guo W, Li X *et al.* Viral lysis alters the optical properties and biological availability of dissolved organic matter derived from *Prochlorococcus Picocyanobacteria*. *Appl Environ Microbiol* 2021;**87**:02271–20. https://doi.org/10.1128/AEM.02271-20.
19. Koonin EV, Makarova KS. CRISPR-Cas: an adaptive immunity system in prokaryotes. *F1000 Biol Rep* 2009;**1**:95. https://doi.org/10.3410/B1-95.
20. Barrangou R, Fremaux C, Deveau H *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;**315**:1709–12. https://doi.org/10.1126/science.1138140.
21. Dion MB, Plante P-L, Zufferey E *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res* 2021;**49**:3127–38. https://doi.org/10.1093/nar/gkab133.
22. Pawluk A, Davidson AR, Maxwell KL. Anti-CRISPR: discovery, mechanism and function. *Nat Rev Microbiol* 2018;**16**:12–7. https://doi.org/10.1038/nrmicro.2017.120.
23. Mendoza SD, Nieweglowska ES, Govindarajan S *et al.* A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* 2020;**577**:244–8. https://doi.org/10.1038/s41586-019-1786-y.
24. Malone LM, Warring SL, Jackson SA *et al.* A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nat Microbiol* 2020;**5**:48–55. https://doi.org/10.1038/s41564-019-0612-5.
25. Borges AL, Davidson AR, Bondy-Denomy J. The discovery, mechanisms, and evolutionary impact of anti-CRISPRs. *Annu Rev Virol* 2017;**4**:37–59. https://doi.org/10.1146/annurev-virology-101416-041616.
26. Marino ND, Pinilla-Redondo R, Csörgő B *et al.* Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. *Nat Methods* 2020;**17**:471–9. https://doi.org/10.1038/s41592-020-0771-6.
27. Bondy-Denomy J, Garcia B, Strum S *et al.* Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature* 2015;**526**:136–9. https://doi.org/10.1038/nature15254.
28. Jia N, Patel DJ. Structure-based functional mechanisms and biotechnology applications of anti-CRISPR proteins. *Nat Rev Mol Cell Biol* 2021;**22**:563–79. https://doi.org/10.1038/s41580-021-00371-9.
29. Burmistrz M, Krakowski K, Krawczyk-Balska A. RNA-targeting CRISPR-Cas systems and their applications. *Int J Mol Sci* 2020;**21**:1122. https://doi.org/10.3390/ijms21031122.
30. Athukoralage JS, McMahon SA, Zhang C *et al.* An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. *Nature* 2020;**577**:572–5. https://doi.org/10.1038/s41586-019-1909-5.
31. Prangishvili D, Bamford DH, Forterre P *et al.* The enigmatic archaeal virosphere. *Nat Rev Microbiol* 2017;**15**:724–39. https://doi.org/10.1038/nrmicro.2017.125.
32. Snyder JC, Bolduc B, Young MJ. 40 years of archaeal virology: expanding viral diversity. *Virology* 2015;**479-480**:369–78. https://doi.org/10.1016/j.virol.2015.03.031.
33. Luk AW, Williams TJ, Erdmann S *et al.* Viruses of haloarchaea. *Life* 2014;**4**:681–715. https://doi.org/10.3390/life4040681.
34. Atanasova NS, Bamford DH, Oksanen HM. Virus-host interplay in high salt environments. *Environ Microbiol Rep* 2016;**8**:431–44. https://doi.org/10.1111/1758-2229.12385.
35. Krupovic M, Cvirkaite-Krupovic V, Iranzo J *et al.* Viruses of archaea: structural, functional, environmental and evolutionary genomics. *Virus Res* 2018;**244**:181–93. https://doi.org/10.1016/j.virusres.2017.11.025.
36. Medvedeva S, Sun J, Yutin N *et al.* Three families of Asgard archaeal viruses identified in metagenome-assembled genomes. *Nat Microbiol* 2022;**7**:962–73. https://doi.org/10.1038/s41564-022-01144-6.

37. Rambo IM, Langwig MV, Leão P *et al.* Genomes of six viruses that infect Asgard archaea from deep-sea sediments. *Nat Microbiol* 2022;**7**:953–61. https://doi.org/10.1038/s41564-022-01150-8.
38. Tamarit D, Caceres EF, Krupovic M *et al.* A closed *Candidatus* Odinarchaeum chromosome exposes Asgard archaeal viruses. *Nat Microbiol* 2022;**7**:948–52. https://doi.org/10.1038/s41564-022-01122-y.
39. Medvedeva S, Borrel G, Krupovic M *et al.* A compendium of viruses from methanogenic archaea reveals their diversity and adaptations to the gut environment. *Nat Microbiol* 2023;**8**:2170–82. https://doi.org/10.1038/s41564-023-01485-w.
40. Laso-Pérez R, Wu F, Crémière A *et al.* Evolutionary diversification of methanotrophic ANME-1 archaea and their expansive virome. *Nat Microbiol* 2023;**8**:231–45. https://doi.org/10.1038/s41564-022-01297-4.
41. Zhou Y, Zhou L, Yan S *et al.* Diverse viruses of marine archaea discovered using metagenomics. *Environ Microbiol* 2023;**25**:367–82. https://doi.org/10.1111/1462-2920.16287.
42. López-Pérez M, Haro-Moreno JM, de la Torre JR *et al.* Novel *Caudovirales* associated with marine group I *Thaumarchaeota* assembled from metagenomes. *Environ Microbiol* 2019;**21**:1980–8. https://doi.org/10.1111/1462-2920.14462.
43. Kim J-G, Kim SJ, Cvirkaite-Krupovic V *et al.* Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea. *Proc Natl Acad Sci U S A* 2019;**116**:15645–50. https://doi.org/10.1073/pnas.1905682116.
44. Ahlgren NA, Fuchsman CA, Rocap G *et al.* Discovery of several novel, widespread, and ecologically distinct marine *Thaumarchaeota* viruses that encode amoC nitrification genes. *ISME J* 2019;**13**:618–31. https://doi.org/10.1038/s41396-018-0289-4.
45. Philosof A, Yutin N, Flores-Uribe J *et al.* Novel abundant oceanic viruses of uncultured marine group II *Euryarchaeota*. *Curr Biol* 2017;**27**:1362–8. https://doi.org/10.1016/j.cub.2017.03.052.
46. Nigro OD, Jungbluth SP, Lin HT *et al.* Viruses in the oceanic basement. *mBio* 2017;**8**:02129–16. https://doi.org/10.1128/mBio.02129-16.
47. Yi Y, Liu S, Hao Y *et al.* A systematic analysis of marine lysogens and proviruses. *Nat Commun* 2023;**14**:6013. https://doi.org/10.1038/s41467-023-41699-4.
48. Krupovic M, Dolja VV, Koonin EV. The LUCA and its complex virome. *Nat. Rev. Microbiol.* 2020;**18**:661–70. https://doi.org/10.1038/s41579-020-0408-x.
49. Li Z, Pan D, Wei G *et al.* Deep sea sediments associated with cold seeps are a subsurface reservoir of viral diversity. *ISME J* 2021;**15**:2366–78. https://doi.org/10.1038/s41396-021-00932-y.
50. Yin X, Cai M, Liu Y *et al.* Subgroup level differences of physiological activities in marine *Lokiarchaeota*. *ISME J* 2021;**15**:848–61. https://doi.org/10.1038/s41396-020-00818-5.
51. Hou J, Wang Y, Zhu P *et al.* Taxonomic and carbon metabolic diversification of Bathyarchaeia during its coevolution history with early earth surface environment. *Sci Adv* 2023;**9**:eadf5069. https://doi.org/10.1126/sciadv.adf5069.
52. Roux S, Páez-Espino D, Chen IMA *et al.* IMG/VR v3: an integrated ecological and evolutionary framework for interrogating genomes of uncultivated viruses. *Nucleic Acids Res* 2021;**49**:D764–75. https://doi.org/10.1093/nar/gkaa946.
53. Koonin EV, Dolja VV, Krupovic M *et al.* Viruses defined by the position of the virosphere within the replicator space. *Microbiol Mol Biol Rev* 2021;**85**:e0019320. https://doi.org/10.1128/MMBR.00193-20.
54. Bin Jang H, Bolduc B, Zablocki O *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* 2019;**37**:632–9. https://doi.org/10.1038/s41587-019-0100-8.
55. Krupovic M, Kuhn JH, Wang F *et al.* Adnaviria: a new realm for archaeal filamentous viruses with linear A-form double-stranded DNA genomes. *J Virol* 2021;**95**:e0067321. https://doi.org/10.1128/JVI.00673-21.
56. Liu Y, Demina TA, Roux S *et al.* Diversity, taxonomy, and evolution of archaeal viruses of the class *Caudoviricetes*. *PLoS Biol* 2021;**19**:e3001442. https://doi.org/10.1371/journal.pbio.3001442.
57. Sheppard C, Werner F. Structure and mechanisms of viral transcription factors in archaea. *Extremophiles* 2017;**21**:829–38. https://doi.org/10.1007/s00792-017-0951-1.
58. Kazlauskas D, Krupovic M, Venclovas Č. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res* 2016;**44**:4551–64. https://doi.org/10.1093/nar/gkw322.
59. Brosius J, Arfsten U. Primary structure of protein L19 from the large subunit of *Escherichia coli* ribosomes. *Biochemistry* 1978;**17**:508–16. https://doi.org/10.1021/bi00596a021.
60. Mizuno CM, Guyomar C, Roux S *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nat Commun* 2019;**10**:752. https://doi.org/10.1038/s41467-019-08672-6.
61. Wang F, Cvirkaite-Krupovic V, Vos M *et al.* Spindle-shaped archaeal viruses evolved from rod-shaped ancestors to package a larger genome. *Cell* 2022;**185**:1297–1307.e11. https://doi.org/10.1016/j.cell.2022.02.019.
62. Quemin Emmanuelle RJ, Pietilä MK, Oksanen HM *et al.* Sulfolobus spindle-shaped virus 1 contains glycosylated capsid proteins, a cellular chromatin protein, and host-derived lipids. *J Virol* 2015;**89**:11681–91. https://doi.org/10.1128/JVI.02270-15.
63. Zhan Z, Zhou J, Huang L. Site-specific recombination by SSV2 integrase: substrate requirement and domain functions. *J Virol* 2015;**89**:10934–44. https://doi.org/10.1128/JVI.01637-15.
64. Albers S-V, Meyer BH. The archaeal cell envelope. *Nat Rev Microbiol* 2011;**9**:414–26. https://doi.org/10.1038/nrmicro2576.
65. Khomyakova MA, Merkel AY, Mamiy DD *et al.* Phenotypic and genomic characterization of *Bathyarchaeum tardum* gen. nov., sp. nov., a cultivated representative of the archaeal class *Bathyarchaeia*. *Front Microbiol* 2023;**14**:1214631. https://doi.org/10.3389/fmicb.2023.1214631.
66. Baquero DP, Contursi P, Piochi M *et al.* New virus isolates from Italian hydrothermal environments underscore the biogeographic pattern in archaeal virus communities. *ISME J* 2020;**14**:1821–33. https://doi.org/10.1038/s41396-020-0653-z.
67. Wang F, Baquero DP, Beltran LC *et al.* Structures of filamentous viruses infecting hyperthermophilic archaea explain DNA stabilization in extreme environments. *Proc Natl Acad Sci U S A* 2020;**117**:19643–52. https://doi.org/10.1073/pnas.2011125117.
68. Quemin ER, Lucas S, Daum B *et al.* First insights into the entry process of hyperthermophilic archaeal viruses. *J Virol* 2013;**87**:13379–85. https://doi.org/10.1128/JVI.02742-13.
69. Prangishvili D, Koonin EV, Krupovic M. Genomics and biology of Rudiviruses, a model for the study of virus-host interactions in archaea. *Biochem Soc Trans* 2013;**41**:443–50. https://doi.org/10.1042/BST20120313.
70. Baquero DP, Liu Y, Wang F *et al.* Structure and assembly of archaeal viruses. *Adv Virus Res* 2020;**108**:127–64.
71. Pinilla-Redondo R, Mayo-Muñoz D, Russel J *et al.* Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Res* 2020;**48**:2000–12. https://doi.org/10.1093/nar/gkz1197.

72. Faure G, Shmakov SA, Yan WX *et al.* CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat Rev Microbiol* 2019;**17**:513–25. https://doi.org/10.1038/s41579-019-0204-7.
73. Moya-Beltrán A, Makarova KS, Acuña LG *et al.* Evolution of type IV CRISPR-Cas systems: insights from CRISPR loci in integrative conjugative elements of *Acidithiobacillia*. *Crispr J* 2021;**4**:656–72. https://doi.org/10.1089/crispr.2021.0051.
74. Al-Shayeb B, Skopintsev P, Soczek KM *et al.* Diverse virus-encoded CRISPR-Cas systems include streamlined genome editors. *Cell* 2022;**185**:4574–4586.e16. https://doi.org/10.1016/j.cell.2022.10.020.
75. Zhou Y, Bravo JPK, Taylor HN *et al.* Structure of a type IV CRISPR-Cas ribonucleoprotein complex. *iScience* 2021;**24**:102201. https://doi.org/10.1016/j.isci.2021.102201.
76. Medvedeva S, Liu Y, Koonin EV *et al.* Virus-borne mini-CRISPR arrays are involved in interviral conflicts. *Nat Commun* 2019;**10**:5204. https://doi.org/10.1038/s41467-019-13205-2.
77. Shiimori M, Garrett SC, Graveley BR *et al.* Cas4 nucleases define the PAM, length, and orientation of DNA fragments integrated at CRISPR loci. *Mol Cell* 2018;**70**:814–824.e6. https://doi.org/10.1016/j.molcel.2018.05.002.
78. Hudaiberdiev S, Shmakov S, Wolf YI *et al.* Phylogenomics of Cas4 family nucleases. *BMC Evol Biol* 2017;**17**:232. https://doi.org/10.1186/s12862-017-1081-1.
79. Zhang Z, Pan S, Liu T *et al.* Cas4 nucleases can effect specific integration of CRISPR spacers. *J Bacteriol* 2019;**201**:00747–18. https://doi.org/10.1128/JB.00747-18.
80. Guo Y, Kragelund BB, White MF *et al.* Functional characterization of a conserved archaeal viral operon revealing single-stranded DNA binding, annealing and nuclease activities. *J Mol Biol* 2015;**427**:2179–91. https://doi.org/10.1016/j.jmb.2015.03.013.
81. Wandera KG, Alkhnbashi OS, Bassett HI *et al.* Anti-CRISPR prediction using deep learning reveals an inhibitor of Cas13b nucleases. *Mol Cell* 2022;**82**:2714–26.e4. https://doi.org/10.1016/j.molcel.2022.05.003.
82. Yang L, Zhang L, Yin P *et al.* Insights into the inhibition of type I-F CRISPR-Cas system by a multifunctional anti-CRISPR protein AcrIF24. *Nat Commun* 2022;**13**:1931. https://doi.org/10.1038/s41467-022-29581-1.
83. Parks DH, Chuvochina M, Rinke C *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022;**50**:D785–94. https://doi.org/10.1093/nar/gkab776.
84. Chaumeil P-A, Mussig AJ, Hugenholtz P *et al.* GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 2022;**38**:5315–6. https://doi.org/10.1093/bioinformatics/btac672.
85. Rozewicki J, Li S, Amada KM *et al.* MAFFT-DASH: integrated protein sequence and structural alignment. *Nucleic Acids Res* 2019;**47**:W5–10. https://doi.org/10.1093/nar/gkz342.
86. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;**25**:1972–3. https://doi.org/10.1093/bioinformatics/btp348.
87. Minh BQ, Schmidt HA, Chernomor O *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020;**37**:1530–4. https://doi.org/10.1093/molbev/msaa015.
88. Parks DH, Imelfort M, Skennerton CT *et al.* CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015;**25**:1043–55. https://doi.org/10.1101/gr.186072.114.
89. Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol Biol* 2019;**1962**:1–14.
90. Parks DH, Chuvochina M, Waite DW *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 2018;**36**:996–1004. https://doi.org/10.1038/nbt.4229.
91. Payne LJ, Todeschini TC, Wu Y *et al.* Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res* 2021;**49**:10868–78. https://doi.org/10.1093/nar/gkab883.
92. Tesson F, Hervé A, Mordret E *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun* 2022;**13**:2561. https://doi.org/10.1038/s41467-022-30269-9.
93. Russel J, Pinilla-Redondo R, Mayo-Muñoz D *et al.* CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. *CRISPR J* 2020;**3**:462–9. https://doi.org/10.1089/crispr.2020.0059.
94. Bland C, Ramsey TL, Sabree F *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 2007;**8**:209. https://doi.org/10.1186/1471-2105-8-209.
95. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015;**12**:59–60. https://doi.org/10.1038/nmeth.3176.
96. Camacho C, Coulouris G, Avagyan V *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421. https://doi.org/10.1186/1471-2105-10-421.
97. Guo J, Bolduc B, Zayed AA *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* 2021;**9**:37. https://doi.org/10.1186/s40168-020-00990-y.
98. Nayfach S, Camargo AP, Schulz F *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 2021;**39**:578–85. https://doi.org/10.1038/s41587-020-00774-7.
99. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 2014;**30**:2068–9. https://doi.org/10.1093/bioinformatics/btu153.
100. Shaffer M, Borton MA, McGivern BB *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res* 2020;**48**:8883–900. https://doi.org/10.1093/nar/gkaa621.
101. Steinegger M, Meier M, Mirdita M *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* 2019;**20**:473. https://doi.org/10.1186/s12859-019-3019-7.
102. Grant JR, Enns E, Marinier E *et al.* Proksee: in-depth characterization and visualization of bacterial genomes. *Nucleic Acids Res* 2023;**51**:W484–92. https://doi.org/10.1093/nar/gkad326.
103. Shannon P, Markiel A, Ozier O *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504. https://doi.org/10.1101/gr.1239303.
104. Nishimura Y, Yoshida T, Kuronishi M *et al.* ViPTree: the viral proteomic tree server. *Bioinformatics* 2017;**33**:2379–80. https://doi.org/10.1093/bioinformatics/btx157.

105. Meier-Kolthoff JP, Göker M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinformatics* 2017;**33**: 3396–404. https://doi.org/10.1093/bioinformatics/btx440.
106. Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
107. Mirdita M, Schütze K, Moriwaki Y *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* 2022;**19**:679–82. https://doi.org/10.1038/s41592-022-01488-1.
108. Hatcher EL, Zhdanov SA, Bao Y *et al.* Virus variation resource - improved response to emergent viral outbreaks. *Nucleic Acids Res* 2017;**45**:D482–d490. https://doi.org/10.1093/nar/gkw1065.
109. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 2004;**32**: 1792–7. https://doi.org/10.1093/nar/gkh340.
110. Kalyaanamoorthy S, Minh BQ, Wong TKF *et al.* ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017;**14**:587–9. https://doi.org/10.1038/nmeth.4285.
111. Xie J, Chen Y, Cai G *et al.* Tree visualization by one table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. *Nucleic Acids Res* 2023;**51**:W587–92. https://doi.org/10.1093/nar/gkad359.
112. Schmartz GP, Hartung A, Hirsch P *et al.* PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res* 2022;**50**:D273–d278. https://doi.org/10.1093/nar/gkab1111.
113. Pourcel C, Touchon M, Villeriot N *et al.* CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res* 2020;**48**:D535–44. https://doi.org/10.1093/nar/gkz915.
114. Bondy-Denomy J, Davidson AR, Doudna JA *et al.* A unified resource for tracking anti-CRISPR names. *Crispr J* 2018;**1**:304–5. https://doi.org/10.1089/crispr.2018.0043.
115. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* 2005;**33**: 2302–9. https://doi.org/10.1093/nar/gki524.
116. Ren M, Feng X, Huang Y *et al.* Phylogenomics suggests oxygen availability as a driving force in *Thaumarchaeota* evolution. *ISME J* 2019;**13**:2150–61. https://doi.org/10.1038/s41396-019-0418-8.
117. Dobson L, Reményi I, Tusnády GE. CCTOP: a consensus constrained TOPology prediction web server. *Nucleic Acids Res* 2015;**43**:W408–12. https://doi.org/10.1093/nar/gkv451.