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, 518060 Shenzhen, China

[2]Archaeal Biology Center, Institute for Advanced Study, Shenzhen University, 518060 Shenzhen, China

[3]Shenzhen Key Laboratory of Marine Microbiome Engineering, Institute for Advanced Study, Shenzhen University, 518060 Shenzhen, China

[4]Department of Ocean Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, China

[5]Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, 75015 Paris, France

[6]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA

Running Head: Virus of Bathyarchaeia

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

Abstract

Bathyarchaeia represent a class of archaea common and abundant in sedimentary ecosystems. The virome of Bathyarchaeia so far has not been characterized. 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 inhibition of host defenses. These findings present an outline of the Bathyarchaeia virome and offer a glimpse into their counter-defense mechanisms.

40 **Main**

41 Bathyarchaeia, formerly the Miscellaneous Crenarchaeotal Group (MCG), is an
42 archaeal class widespread in marine and freshwater sediments[1, 2, 3, 4, 5]. The estimated
43 global abundance of Bathyarchaeia reaches up to 2.0-3.9 × 10^{28} cells, representing one
44 of the most abundant groups of archaea on Earth[6]. Genomic analyses suggest that
45 Bathyarchaeia lead an acetyl-CoA-centered heterotrophic lifestyle with the potential
46 for acetogenesis[6], methane metabolism[7], and sulfur reduction[8]. Bathyarchaeia also
47 encompass a variety of genes encoding carbohydrate-active enzymes[9] and thus likely
48 can unitize various carbohydrates and lignin[10]. The diverse metabolic potential of
49 Bathyarchaeia contributes to their predominance in sedimentary environments,
50 rendering them essential players in the global carbon cycle[5, 8, 11].

51 Viruses, as the most abundant biological agents on the planet[12, 13], reshape the
52 microbial communities through cell lysis[14, 15, 16, 17] and alter host metabolism via virus-
53 encoded auxiliary metabolic genes (AMGs) [18, 19, 20]. Bacteria and archaea evolved
54 enormously versatile repertoires of antivirus immune systems. In particular, nearly all
55 archaea and many bacteria encode CRISPR-Cas, the prokaryotic adaptive immunity
56 systems[21]. The CRISPR-Cas system selectively acquires foreign DNA fragments
57 (protospacers) and stores them as spacers in the CRISPR array which is expressed to
58 produce CRISPR (cr) RNAs that serve as guides recognizing the DNA or RNA target
59 and recruiting CRISPR effector nucleases[22]. The CRISPR-Cas systems provide the
60 most reliable basis for viral host prediction by linking host spacers to the cognate virus
61 protospacers[23].

62 In response to the host defenses, viruses infecting bacteria and archaea evolved a
63 broad repertoire of counter-defense mechanisms to evade immunity[24, 25, 26], engaging in
64 the perennial arms race. In particular, many viruses encode diverse anti-CRISPR (Acr)
65 proteins that specifically target different CRISPR-Cas subtypes[27, 28].
66 To date, only a limited number of archaeal viruses have been isolated by traditional
67 cultivation-dependent methods. In recent years, an increasing diversity of archaeal
68 viruses has been uncovered through metagenomic data mining[29], including viruses
69 associated with Asgard archaea[30, 31, 32], methanotrophic ANME-1 archaea[33], ammonia-
70 oxidizing archaea[34, 35, 36, 37] and marine group II Euryarchaeota (Poseidoniales)[38].
71 However, no viruses linked to Bathyarchaeia, the most prominent group of archaea in
72 the sedimentary ecosystems, have been reported. Here we describe the results of
73 metagenomic analysis revealing a substantial diversity of viruses associated with
74 Bathyarchaeia, some of which encode CRISPR-Cas system components and predicted
75 Acr.
76
77 **Discovery of viruses associated with Bathyarchaeia**
78 In total, 367 metagenome-assembled genomes (MAGs) of Bathyarchaeia,
79 including 68 high-quality MAGs (Extended Data Table S1), were obtained from our
80 previous results[8, 39] and public databases. Phylogenetic analysis based on a
81 concatenated alignment of an optimized set of 51 marker proteins (see Methods)
82 yielded 8 major clades of Bathyarchaeia that correspond to the orders in the latest
83 taxonomy[40] (Fig. 1a and Extended Data Fig. S1).

We then searched for CRISPR-Cas systems in Bathyarchaeia MAGs and identified type I and type III CRISPR-*cas* loci in over one-fifth of the MAGs. Typically, a bathyarchaeal genome carries a single CRISPR-Cas system of, either type I (53 high quality MAGs) or type III (15 high quality MAGs). Notably, however, 11 MAGs were found to encompass both types 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 Methods), 49 high-confidence CRISPR array from different Bathyarchaeia orders were identified, containing 1,602 spacers (Extended Data Fig. S2). The spacers were used to search for protospacers in the viral sequences of IMG/VR v3 database[41]. Additionally, a search for potential proviruses in the host genomes was conducted (see 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[42]. Six viral contigs corresponded to complete genomes, as indicated by the presence of terminal repeats (Extended Data 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 6 viral clusters (VCs) (Fig. 2). Additionally, we carried out a genome-wide sequence similarity comparison and examined the phylogenies of hallmark genes in Bathyarchaeia viruses, comparing them with known archaeal viruses (Extended Data Fig. S3). Based on the results of these

analyses, we identified four putative family-level groups of Bathyarchaeia viruses. The families '*Fuxiviridae*' and '*Kunpengviridae*' include head and tail viruses of the class *Caudoviricetes* in the realm *Duplodnaviria*. The family '*Chiyouviridae*' consists of filamentous viruses of the archaea-specific realm *Adnaviria*[43]. 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. 2).

The proposed family '*Fuxiviridae*' is represented by three nearly identical complete genomes (Fuxivirus) that encompass protospacers targeted by type I-A CRISPR spacers from the Bathyarchaeia order Bifangarchaeales (Fig. 1 and Fig. 3). 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. 3a). Fuxivirus encodes all the hallmark proteins of *Caudoviricetes*, namely, a HK97-like MCP (Extended Data Fig. S3b and Fuxivirus_34 in Extended Data Table S4), a portal protein (Fuxivirus_28), a terminase large subunit (Fuxivirus_27), a tail tube protein (Fuxivirus_38) and several other tail components (Fig. 3a and Extended Data Table S4) which are similar to previously characterized archaeal tailed viruses[44]. 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. 3a and Extended Data Table S4) that likely regulate viral gene expression[45]. Notably, Fuxivirus also encodes CRISPR-Cas system components, namely, a type IV *cas* gene cluster (Fuxivirus_1 to Fuxivirus_5 in Extended Data Table

128 S4) with a mini CRISPR array, and a Cas4 protein (Fuxivirus_12) (Fig. 3a and see
129 below).

130 The proposed family '*Kunpengviridae*' includes one complete viral genome
131 (Kunpengvirus), that is targeted by a single spacer (100% match) from *Bathyarchaeia*
132 sp. QMXD of the order Jinwuousiales (Fig. 1). (Fig. 3b). Kunpengvirus encodes the
133 hallmark proteins of *Caudoviricetes*, as well as a suite of tail proteins including the
134 baseplate protein (Fig. 3b and Extended Data Table S4). Kunpengvirus also encodes an
135 integrase (Kenpengvirus_28), suggesting that it can integrate into the host genome as a
136 provirus. In addition, this virus encompasses genes for a deoxynucleoside
137 monophosphate kinase (dNMPK) (Kenpengvirus_41), an MCM-like helicase
138 (Kenpengvirus_46) and a family B DNA polymerase (Kenpengvirus_42) (Fig. 3b and
139 Extended Data Table S4) which are indicative of (at least, partially) autonomous
140 genome replication[46]. Notably, unlike in other tailed viruses, the proofreading DEDDy
141 3'-5' exonuclease (Kenpengvirus_45) and the family B DNA polymerase are encoded
142 by two distinct genes. Additionally, Kunpengvirus encodes a homolog of ribosomal
143 protein bL19 (Kunpengvirus_2 in Extended Data Table S4), which is typically present
144 in bacteria and eukaryotes (chloroplasts and mitochondria) but not in archaea, except
145 for *Candidatus* Aenigmarchaeota. The bL19 is located at the 30S-50S ribosomal
146 subunit interface and is thought to contribute to the structure and function of the
147 aminoacyl-tRNA binding site[47]. A number of ribosomal proteins, including bL19, were
148 previously identified in bacterial viruses[48], but the only ribosomal protein so far
149 detected in archaeal viruses had been L21e[44]. The presence of bL19 in Kunpengvirus

150 suggests that modification of the host translation apparatus by archaeal viruses is more
151 common than currently recognized. Similar to the Fuxivirus, Kunpengvirus also
152 encodes a Cas4 endonuclease (Fig. 3b and Kenpengvirus_47 in Extended Data Table
153 S4).

154 A unique viral genome (Huangdivirus) representing the proposed family
155 '*Huangdiviridae*' was identified as an apparent provirus in *Bathyarchaeia* sp. QMYA
156 of the order Baizomonadales from a deep-sea hydrothermal vent (Fig. 1). Sensitive
157 sequence comparison using HHsearch identified three virus-encoded structural proteins
158 (VPs) VP1-3 (Fig. 3c and Extended Data Table S4), homologous to the structural
159 proteins of archaeal spindle-shaped viruses of the family *Fuselloviridae*, as well as an
160 AAA+ ATPase (Huangdivirus_17, HHblits best hit to ATV ATPase, with 99.8%
161 probability). Structural predictions for VP1 (Huangdivirus_21) and VP3
162 (Huangdivirus_28) indicated that both proteins contain two hydrophobic α-helices
163 connected by a short turn (Extended Data Fig. S5a) resembling the typical structure of
164 the MCPs of spindle-shaped viruses[49]. Huangdivirus VP2 (Huangdivirus_19) is most
165 closely related to the viral DNA-binding protein VP2 of Sulfolobus spindle-shaped
166 virus (SSV1)[50] (HHblits probability of 99.49%). As in the case of SSV1[50], we detected
167 the consensus glycosylation motifs (N-X-S/T) in Huangdivirus VP1 and VP3 which
168 may be glycosylated by the virus-encoded glycosyltransferase (Huangdivirus_40 and
169 Extended Data Fig. S4b). Capsid protein glycosylation is thought to contribute to the
170 virion stability, particularly for viruses infecting hyperthermophilic hosts[50]. Similar to
171 fuselloviruses infecting hyperthermophilic archaea of the order Sulfolobales[51],

172 Huangdivirus encodes a putative integrase of the tyrosine recombinase superfamily (Fig.
173 3c and Huangdivirus_16 in Extended Data Table S4) which is likely to be responsible
174 for viral DNA integration into the host chromosome. Notably, Huangdivirus encodes a
175 predicted polysaccharide lyase (Fig.3c and Huangdivirus_33 in Extended Data Table
176 S4) that might be involved in the break-down of the S-layer polysaccharides[52, 53] of host
177 cell wall during infection.

178 '*Chiyouviridae*' is a potential new family of filamentous viruses in the order
179 *Ligamenvirale* (class *Tokiviricetes*, realm *Adnaviria*). '*Chiyouviridae*' is represented by
180 one complete viral genome (Chiyouvirus) that is targeted by two spacers of
181 *Bathyarchaeia* sp. JAGTQC in the order Bifangarchaeales (Fig. 1 and Fig. 3d).
182 Phylogenomic analysis of all available *Tokiviricetes*[54] genomes recapitulated the
183 previously established relationships and showed that Chiyouvirus forms a separate
184 clade within the order *Ligamenvirales*, most closely related to the families *Rudiviridae*
185 and *Ungulaviridae* (Extended Data Fig. S6a). Whole proteome comparison showed less
186 than 50% average amino acid identity (AAI) between protein homologs from
187 Chiyouvirus and members of other viral families, with the highest AAI (46%) with the
188 genus *Icerudivirus* of the family *Rudiviridae* (Extended Data Fig. S6b). Similar to all
189 members of the families *Ungulaviridae* and *Lipothrixviridae* but only some members
190 of the family *Rudiviridae*[54], Chiyouvirus encodes two MCPs (Chiyouvirus_14 and
191 Chiyouvirus_23), each comprising an alpha-helix bundle (Fig. 3d and Extended Data
192 Fig. S6c). In rudiviruses, unlike other members of the realm *Adnaviria*, the second MCP
193 paralog, when present, is not incorporated into virions[55]. Thus, it remains unclear

194 whether both Chiyouvirus MCPs are involved in virion formation. In addition,
195 Chiyouvirus encodes a large minor structural protein (Fig. 3d and Chiyouvirus_18 in
196 Extended Data Table S4), a homolog of SIRV2 P1070, which is thought to be involved
197 in the formation of the virion terminal filaments that are responsible for host
198 recognition[56, 57]. Predictions for transmembrane proteins indicated that Chiyouvirus
199 encodes 8 potential transmembrane proteins (Fig. 3d) suggesting that it is a membrane-
200 enveloped filamentous virus, similar to the adnaviruses in the families *Lipothrixviridae*,
201 *Ungulaviridae* and *Tristromaviridae*[58].
202

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

205 We explored the type IV CRISPR-Cas system encoded by '*Fuxiviridae*' in greater
206 detail, including phylogenetic analysis and gene locus comparison. The viral CRISPR-
207 *cas* locus encodes all typical components of the type IV CRISPR-Cas effector module
208 (Fig. 3a and Extended Data Table S4), in particular, the signature protein Csf1
209 (Fuxivirus_5 in Extended Data Table S4), the apparent large subunit of the effector
210 complex, but no adaptation module. Phylogenetic analysis of Cas proteins, including
211 Csf2 (Cas7) (Fuxivirus_3 in Extended Data Table S4), the most conserved protein in
212 the type IV systems[59], showed that CRISPR-Cas system of '*Fuxiviridae*' belongs to the
213 IV-B subtype (Extended Data Fig. S7). Additionally, a CysH-like protein (Fuxivirus_1
214 in Extended Data Table S4), which is tightly associated with the type IV-B systems, is
215 encoded in the virus genome adjacent to Csf3 (Cas5) (Fuxivirus_2; Fig. 3a).

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 Fig. S7c). Previously, type IV CRISPR-Cas systems have been observed in plasmids and prophages[60, 61] as well as several lytic phages[62], 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 genomes contains a CRISPR mini-array that consists of two repeats and a spacer (Fig. 3a). 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[63]. However, given the presence of the mini-array, we hypothesize 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[64]. 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.

The inferred Fuxivirus host genome, *Bathyarchaeia* sp. DRVA, contains two CRISPR arrays and two *cas3* gene but lacks a complete CRISPR-Cas system. Another genome of *Bathyarchaeia* sp. JAGTQM, from the same order Bifangarchaeales, was found to share identical CRISPR repeat and a closely similar Cas3 protein (84.03%

238 identity, 100% coverage) with *Bathyarchaeia* sp. DRVA. However, *Bathyarchaeia* sp.
239 JAGTQM contains complete type I-A and III-D CRISPR-Cas system, suggesting the
240 presence of a CRISPR-Cas system(s) in the DRVA genome as well. Notably, we found
241 that the *cas7* gene of Fuxivirus was targeted by a host spacer (Fig. 3a), suggesting that
242 the virus-encoded type IV-B system, and with it, possibly, the virus reproduction, can
243 be inhibited by the host CRISPR-Cas systems.

244 The Cas4 homolog encoded by Fuxivirus could be an additional counter-defense
245 factor (Fig. 3a and Fuxivirus_12 in Extended Data Table S4). Cas4 is a P-D/ExK family
246 nuclease that is a common component of CRISPR-Cas systems that assists Cas1-Cas2
247 integration complexes in the acquisition of CRISPR spacers[65], but many Cas4
248 homologs are encoded outside CRISPR-*cas* loci[66]. The Cas4 homologs are encoded by
249 both '*Fuxiviridae*' (Fuxivirus_12) and '*Kunpengviridae*' (Kenpengvirus_47).
250 Phylogenetic analysis showed that these Bathyarchaeia viral Cas4 proteins are most
251 closely related to Cas4 homologs encoded by Sulfolobus-infecting rudiviruses
252 (Extended Data Fig. S7b), suggesting the possibility of horizontal gene transfer
253 between unrelated archaeal viruses. Multiple sequence alignment confirmed that two
254 Cas4 homologs encoded by '*Fuxiviridae*' and '*Kunpengviridae*' share nearly all
255 conserved amino acids with the rudivirus SIRV2-encoded Cas4 homolog and are thus
256 predicted to be active nucleases (Extended Data Fig. S8). Previous studies have
257 demonstrated that overexpression of the SIRV2-encoded Cas4 in the archaeal host
258 substantially reduced the efficiency of exogenous spacers acquisition by the host
259 CRISPR-Cas system[67]. 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[68].

We identified two mini-arrays in Chiyouvirus genome, each containing 4 repeats and 3 spacers (Fig. 3d), 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[60]. Furthermore, the virus could potentially recruit the host CRISPR adaptation machinery to incorporate 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[60, 64], 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. 3d). 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 (Chiyouvirus_39 in Extended Data Table S4), with a conserved Mg^{2+} binding site (Extended Data Fig. S9), 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[69]. We found that the structural model of a predicted Chiyouvirus Acr protein (Chiyouvirus_17 in Extended Data Table

282 S4) was significantly similar to the N-terminal domain of AcrIF24 (7DTR, chain A)
283 (Extended Data Fig. S10) which inhibits the activity of a type I-F CRISPR-Cas system[70]
284 and thus could function as an Acr as well.

285 Finally, in addition to the components of CRISPR-Cas systems, Fuxivirus and
286 Kunpengvirus encode stand-alone DNA methyltransferases (Fuxivirus_47 and
287 Kenpengvirus_43 in Extended Data Table S4) that could provide protection against the
288 host restriction-modification systems.

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

295

296 **Methods**

297 **Bathyarchaeia genome collection and classification.** Bathyarchaeia MAGs
298 (metagenome-assembled genomes) were obtained from our previous dataset[8, 39] and the
299 NCBI (https://www.ncbi.nlm.nih.gov) and IMG/M (https://jgi.doe.gov) public database
300 (up until June 30th, 2021). Bathyarchaeia genome classification was based on 51 high-
301 coverage marker gene optimized from GTDB[71] v207 (TIGR01171 and TIGR02389
302 were excluded due to the low coverage). Marker genes were initially identified with
303 GTDB-Tk v2.1.1[72] using 'identify' method. Only MAGs contain more than 50% of the

304 marker genes were kept. Sequences were then aligned using MAFFT v7.457[73] and
305 trimmed with TrimAl v1.4[74] (-gappyout). The maximum likelihood phylogenomic tree
306 for concatenated 51 proteins was constructed with IQ-TREE 2.0.6 (best-fit model
307 LG+F+R13, -B 1000, -alrt 1000). The quality, contamination, GC content, and other
308 sequence information of Bathyarchaeia MAGs were assessed by checkM v1.2.2[75].
309 Biome information was extracted from the GenBank file or relevant literatures.
310 To delineate Bathyarchaeia orders, the maximum likelihood tree based on 51 proteins
311 was converted to an ultra-metric tree and the relative evolutionary divergence
312 (RED)[76] was calculated. An order was called when its branch length in the ultrametric
313 tree corresponded to a RED value within the range of 0.5 to 0.7.
314
315 **CRISPR-Cas system detection and CRISPR array validation.** CRISPR-Cas
316 systems in Bathyarchaeia MAGs were detected using DefenseFinder[77] and further
317 classified using CRISPRCasTyper[78]. CRISPR array in Bathyarchaeia MAGs were
318 searched with MinCED[79] (Default setting). To remove ambiguous CRISPR array that
319 might not belong to Bathyarchaeia, all CRISPR array-containing contigs encoding
320 proteins were searched against the nr database (2022-09) using Diamond (2.1.6)
321 BLASTP[80] (-evalue 1e-5, --more-sensitive), and only contigs with at least 3 proteins
322 assigned to Bathyarchaeia were retained. The validated CRISPR repeats were clustered
323 with 90% identity using an all-against-all BLASTN[81] search (E-value cutoff 1e-5, word
324 size 7).
325

326 **Bathyarchaeia virus identification and annotation.** We first carried out the CRISPR-
327 based virus-host assignments. Spacers from validated CRISPR array were used as a
328 nucleotide database to compare to viral contigs in the IMG/VR v3[41] database using
329 BLASTN[81] with parameters (-task blastn-short -evalue 1e-5). Only viral contigs
330 harboring protospacers with 100% coverage and a maximum of one mismatch were
331 considered as Bathyarchaeia viruses.
332 In addition, some viruses might form proviruses that cannot be detected by CRISPR-
333 protospacer search. Therefore, we used VirSorter2 v.2.2.3[82] to identified potential viral
334 sequences in Bathyarchaeia MAGs. Specifically, all the Bathyarchaeia contigs were
335 filtered using VirSorter2 (--include-groups dsDNAphage, ssDNA --min-length 5000 --
336 min-score 0.5) and only contigs with identifiable viral structural proteins and at least
337 one ORF identified through a BLASTP search against the nr database as Bathyarchaeia
338 were retained. Virial genome completeness estimation and the potential host regions
339 trimming were conducted by CheckV 0.8.1[83]. Coding sequences were identified using
340 Prokka[84] with parameters (--kingdom viruses --gcode 1).
341 All the Bathyarchaeia viral proteins identified as describe above were annotated with
342 DRAM[85] (viral mode) and HHblits[86] (MSA generated with UniRef30_2020_06 with 3
343 interactions, Evalue 1e-6, MSA was compared against the PDB_mmCIF70_14_Apr,
344 SCOPe70_2.07, Pfam-A_v35 and UniProt-SwissProt-viral70_3_Nov_2021 databases
345 with parameters -Z 250 -loc -z 1 -b 1 -B 250 -ssm 2 -sc 1 -seq 1 -dbstrlen 10000 -
346 norealign -maxres 32000). All the viral genome maps were visualized using Proksee[87].
347

348 **Bathyarchaeia viral taxonomy assignment.** To determine the taxonomic status of
349 Bathyarchaeia viruses, the whole genome network analysis was carried out with
350 vConTACT2[88] (default parameters) against the Viral RefSeq v207 database as well as
351 archaeal viruses proposed in the latest International Committee on Taxonomy of
352 Viruses Report (VMR_MSL38_v1). The resulting networks were visualized using
353 Cytoscape v3.9.1[89] with an edge-weighted spring embedded model.
354 For viruses in the realm *Duplodnaviria*, proteome-scale phylogeny was constructed
355 using the VipTree version 3.4[90]. The analysis was carried out with archaea viral families
356 in the realm *Duplodnaviria*, according to the latest ICTV taxonomy classification
357 (04/25/2023). For filamentous viruses, proteome-scale phylogeny was constructed by
358 using VICTOR[91] with all known members of the *Tokiviricetes* class. Structural
359 modeling of viral proteins was performed with AlphaFold2[92] via ColabFold v1.5.1[93]
360 with pdb70 template mode.
361
362 **Bathyarchaeia viral CRISPR-Cas systems classification.** All the Cas proteins were
363 manually search with HHblits (see above). To determine the CRISPR-Cas system
364 subtype of '*Fuxiviridae*', phylogenetic tree of all the Cas proteins were constructed with
365 reference sequences. Specifically, Cas4 proteins were aligned with 29 reference[67]
366 sequences by MUSCLE 5.1[94] and trimmed by TrimAl v1.4 (-gappyout). The maximum
367 likelihood tree was inferred by IQ-TREE 2.0.6 (-m LG+I+G4, -B 1000, -alrt 1000).
368 Cas7 proteins were aligned with 296 reference sequences[95] by MUSCLE 5.1. The
369 alignment was trimmed by TrimAl v1.4 (-gt 0.7) and sequence with more than 50%

370 gaps were dropped. The maximum likelihood tree was inferred by IQ-TREE 2.0.6 (-m
371 VT+R10, -B 1000 -alrt 1000). Cas5 proteins were aligned with 722 reference[95]
372 sequences by MUSCLE 5.1. The alignment was trimmed by TrimAl v1.4 (-gappyout)
373 and sequence with more than 50% gaps were dropped. The maximum likelihood tree
374 was inferred by IQ-TREE 2.0.6 (-m VT+R10, -B 1000 -alrt 1000). Cas11 proteins were
375 aligned with 45 reference sequences[59, 95] by MUSCLE 5.1. The alignment was trimmed
376 by TrimAl v1.4 (-gappyout) and sequence with more than 50% gaps were dropped. The
377 maximum likelihood tree was inferred by IQ-TREE 2.0.6 (-m VT+R3, -B 1000 -alrt
378 1000). Csf1 proteins were aligned with 49 reference sequences[59, 95] sequences by
379 MUSCLE 5.1. The alignment was trimmed by TrimAl v1.4 (-gappyout) and sequence
380 with more than 50% gaps were dropped. The maximum likelihood tree was inferred by
381 IQ-TREE 2.0.6 (-m WAG+I+G4, -B 1000 -alrt 1000). Phosphoadenosine
382 phosphosulfate (PAPS) reductase (CysH) domain containing protein has been observed
383 in prokaryotes and virus[61]. We first recalled all the CysH-like proteins from 18988
384 public reference complete viral genomes[96]. Together with *bona fide* CysH and Cas-
385 associated CysH-like proteins, 322 sequences were aligned by MUSCLE 5.1 and
386 trimmed by TrimAl v1.4 (-gappyout). The maximum likelihood tree was inferred by
387 IQ-TREE 2.0.6 (-m WAG+F+R7, -B 1000 -alrt 1000).
388 All final phylogenetic trees were visualized by tvBOT[97].
389
390 **Bathyarchaeia viral CRISPR spacer-protospacer match analysis**
391 CRISPR array in Bathyarchaeia viral genome were detected with MinCED (-minNR 1).

392 All the spacer obtained from Bathyarchaeia virus were BLASTN search (-task blastn-
393 short) with viral database IMG/VR v3 and plasmid sequences database PLSDB
394 database[98] (v. 2021_06_23_v2) as well as Bathyarchaeia MAGs in this study. Direct
395 repeats were searched against CRISPR-CAS++ databases[99] and Bathyarchaeia CRISPR
396 repeat dataset with BLASTN (-evalue 0.01).
397

398 **Bathyarchaeia viral anti-CRISPRs proteins (Acrs) identification**

399 The potential anti-CRISPR proteins (Acrs) of Bathyarchaeia viruses were first
400 predicted using DeepAcr[69]. The protein structures of predicted Acrs, along with Acrs
401 from reference[100], were modeled using AlphaFold2 by ColabFold v1.5.2 in pdb70
402 template mode. These potential Bathyarchaeia virus Acrs were then used as query
403 models for alignment to the reference Acr models using TM-align[101], and only models
404 with a TM score greater than 0.3 were retained. This initial computational analysis was
405 further supplemented with comprehensive meticulous manual inspection of the
406 alignments and structural models.
407

408 **Etymology**

409 '*Fuxiviridae*': Derived from Fuxi, a legendary figure in Chinese mythology known
410 for his diverse talents and abilities. This alludes to the possibility that the virus has a
411 multifaceted counter-defense systems, capable of employing various strategies to evade
412 the host's immune response.
413 '*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.

Acknowledgments

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 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.

This work was supported by grants from the 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). E.V.K. is supported by the Intramural Research

Program of the National Insitutes of Health of the USA (National Library of Medicine).

Author contributions

M.L. and C.D. designed the experiments; C.D., Yang L., Ying L. and M.K. provided support on bioinformatic analysis. C.D. performed the data analyses and wrote the paper with contributions from all the authors.

Competing interests

The authors declare that they have no competing interests.

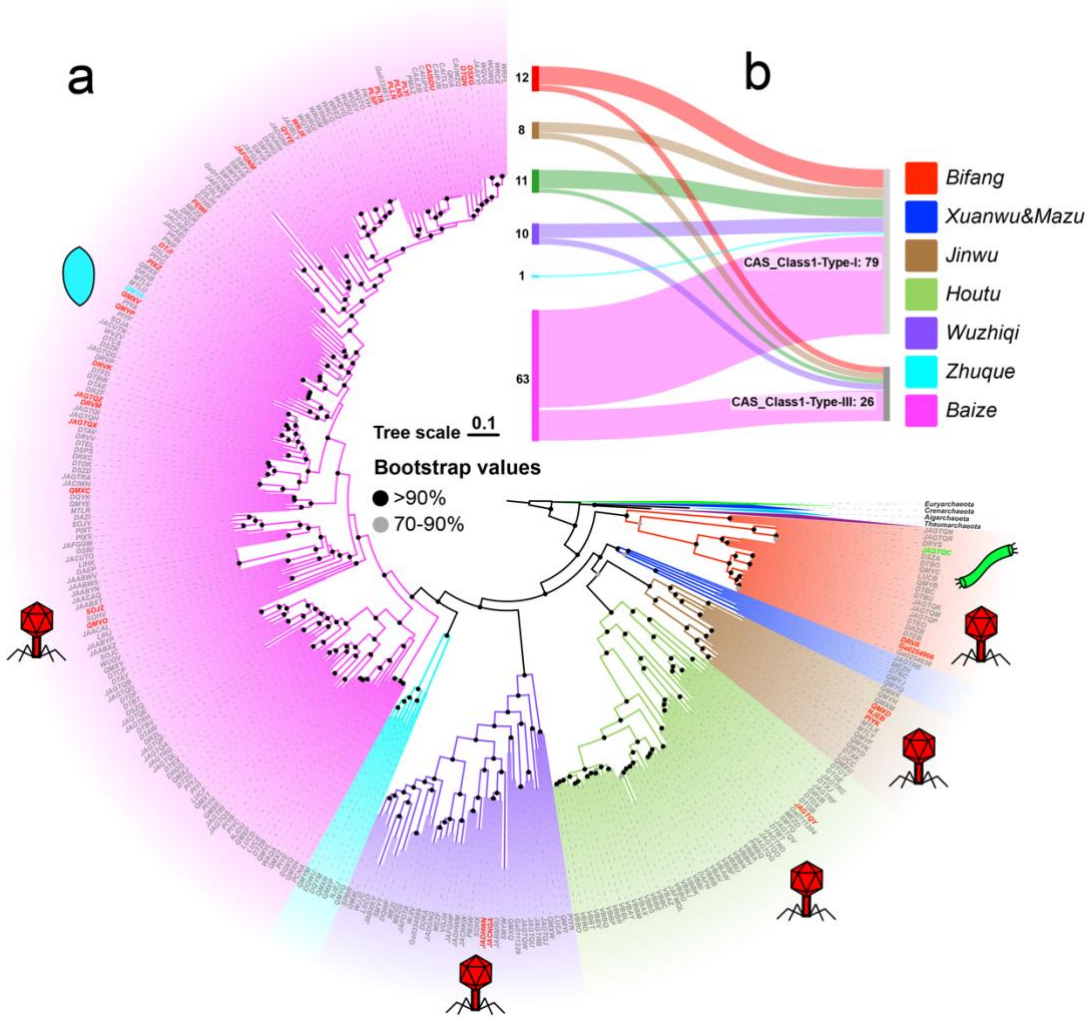


Fig. 1 Distribution of identified viruses across the evolutionary tree of the Bathyarchaeia. a) Maximum likelihood tree of Bathyarchaeia was reconstructed based on the modified set of 51 marker genes. The outgroup was taken from reference[102] and includes representatives of the phyla Euryarchaeota, 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; Xuanwu, Xuanwuarculales; Jinwu, Jinwuousiales; Mazu, Mazuousiales; Houtu, Houtuarculales; Wuzhiqi, Wuzhiqiibiales; Zhuque, Zhuquarculales; Baize, Baizomonadales.

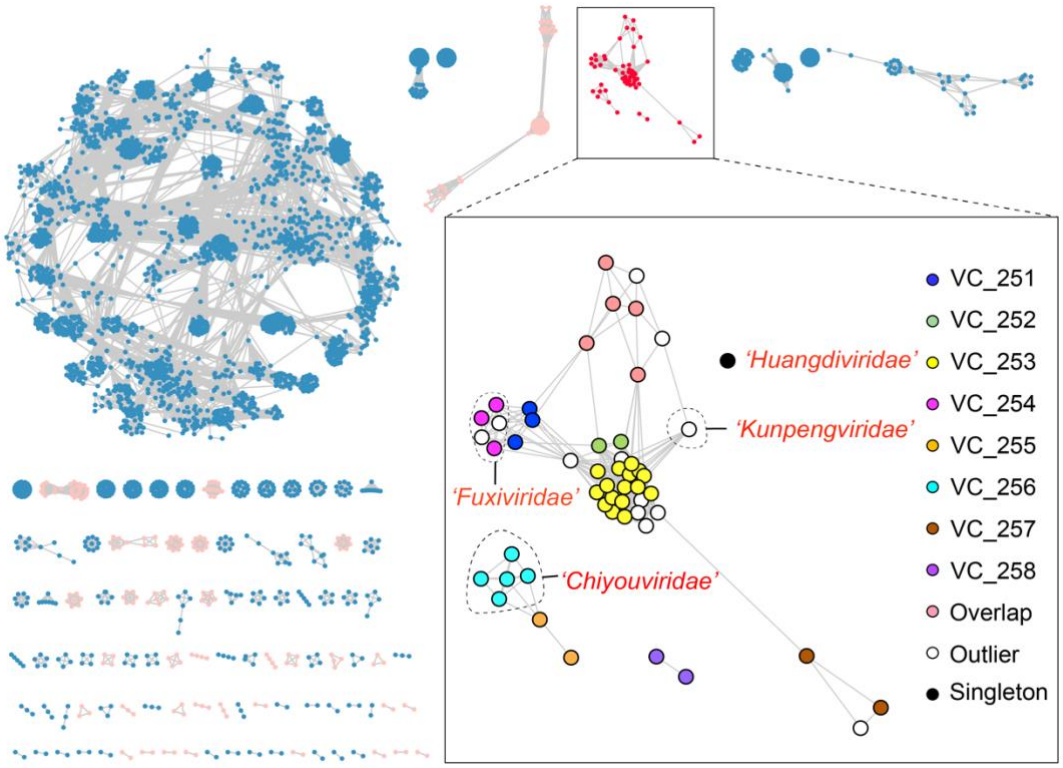


Fig. 2 Classification of Bathyarchaeia viruses based on the whole-genome protein-sharing network with other prokaryotic viruses. The whole-genome protein-sharing network analysis was constructed using vConTACT2 v.0.11.3 for the taxonomic assignment of 56 Bathyarchaeia viral genomes. Bathyarchaeia viruses clusters are highlighted in red in the complete network. Bathyarchaeia viruses are assigned to 4 distinct groups, including one large cluster. Viral clusters (VCs) are indicated with differently color circles. The proposed virus families including complete genomes are separated by dashed lines and appended with the corresponding names in red. Archaeal viruses are in pink, and bacterial viruses are in blue. The networks were visualized with Cytoscape v.3.9.1.

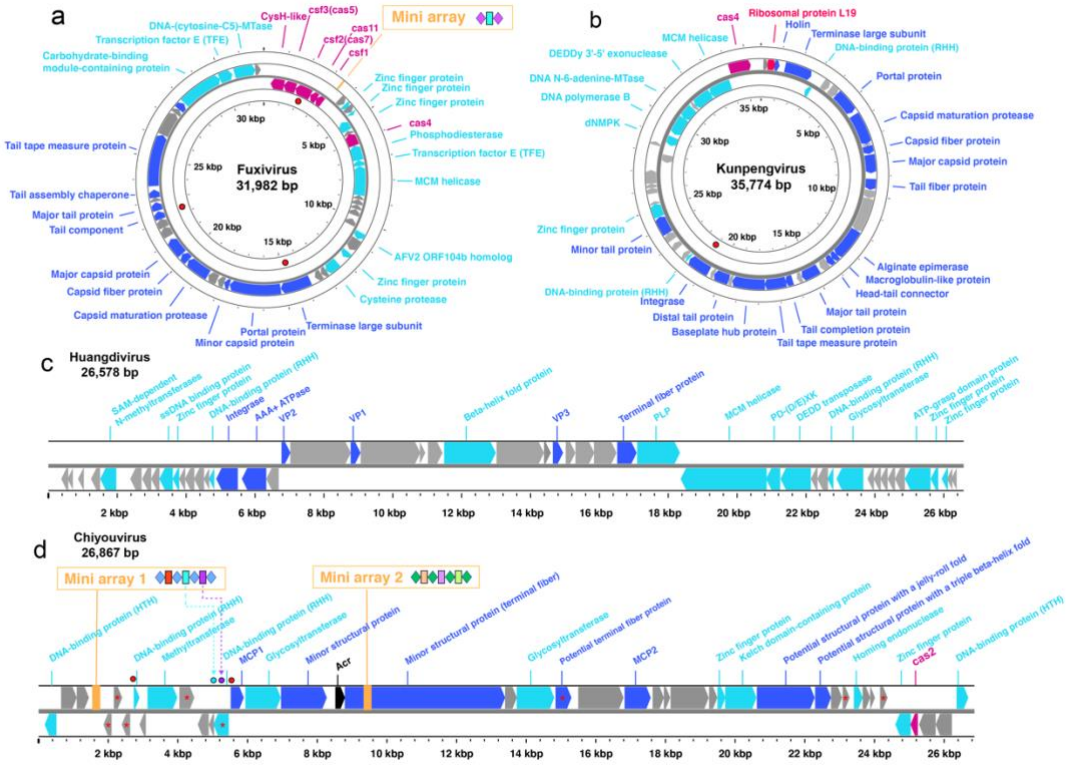


Fig. 3 Maps of complete genomes of Bathyarchaeia viruses. a) Genome map of Fuxivirus. **b**) Genome map of Kunpengvirus. **c**) Genome map of Huangdivirus. **d**) Genome map of Chiyouvirus. Genes annotated by HHblits with a probability greater than 95% are shown as different colors. Genes related to 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[103] server are indicated with red asterisks. Detailed gene annotations are in Supplementary table S4.

Extended Data

Extended Data Fig. S1: The calculated relative evolutionary divergence (RED) value

488 of each node at order level. Eight order-level units of class Bathyarchaeia are marked
489 in different colors.
490
491 Extended Data Fig. S2: Bathyarchaeia CRISPR Array Overview. a) CRISPR repeats
492 found in Bathyarchaeia MAGs, clustered at 90% sequence identity. Verified repeats are
493 highlight in green, while repeats associated with detected viruses in this study are
494 indicated in red. b) Seq-logo for repeat clusters linked to viruses identified in this
495 research. c) Bathyarchaeia CRISPR spacer length distribution.
496
497 Extended Data Fig. S3: The genome-wide sequence similarity comparison,
498 phylogenetic and modeling of major capsid of Bathyarchaeia viruses in realm
499 *Duplodnaviria*. a) The proteomic tree displays the relationship between Bathyarchaeia
500 viruses and archaea viruses in the realm *Duplodnaviria*, based on genome-wide
501 sequence similarities. Bathyarchaeia virus families are labeled in green, and viruses
502 with complete genomes are marked with a red circle. b) The maximum likelihood tree
503 of major capsid proteins (MCPs) shows the relationship between Bathyarchaeia virus
504 and archaeal virus. Bathyarchaeia virus families are labeled in green, and the 3D
505 structure representing the MCP of each family is displayed next to it.
506
507 Extended Data Fig. S4: Reads coverage of Fuxivirus genome. Raw reads from data
508 (SRR5248875) are mapped to the virus genome with 100% identity using Bowtie 2 and
509 Samtools v1.1.7.

510

511 Extended Data Fig. S5: Huangdivirus structural proteins. a) Structural modeling of
512 Huangdivirus capsid proteins VP1 and VP3, colored using a rainbow gradient from N-
513 terminus (blue) to C-terminus (red). b) Sequence analysis of Huangdivirus structural
514 proteins VP1-3, with predicted transmembrane domains highlighted in yellow and
515 theoretical glycosylation consensus motifs (N-X-S/T) shown on a green background.

516

517 Extended Data Fig. S6: Genomic analysis of Chiyouvirus. a) Phylogenomic tree of
518 Chiyouvirus (red arrow) alongside known members of the *Tokiviricetes* class, based on
519 whole-genome amino acid analysis using VICTOR. Tree is rooted with *Primavirales*,
520 and branch length represents GBDP distance formula D6. Branch support values are
521 indicated with numbers. b) Whole-genome amino acid identity comparison of
522 filamentous viruses in the *Tokiviricetes* class, conducted by EzAAI[104]. Chiyouvirus is
523 highlighted in red. Only AAI values greater than 40% are displayed in the heatmap. c)
524 Predicted structural model comparison of Bathyarchaeial Chiyouvirus major capsid
525 proteins MCP1 and MCP2 with Icerudivirus SIRV (3J9X, chain A) and Captovirus
526 AFV1 (5W7G, chain A) structures. Models are colored using a rainbow gradient from
527 N-terminus (blue) to C-terminus (red).

528

529 Extended Data Fig. S7: Phylogenetic analysis of proteins in Fuxivirus type IV-B
530 CRISPR-Cas system and Kunpengvirus Cas4 protein. Maximum likelihood trees for a)
531 Csf2 (Cas7), b) Cas4 c) CysH-like protein, d) Csf3 (Cas5), e) Csf1, and f) Csf4 (Cas11).

Fuxivirus proteins are marked with red pentagrams. Kunpengvirus Cas4 protein is highlighted with blue in figure b).

Extended Data Fig. S8: Sequences alignment of Cas4 protein. Conserved amino acid residues involved in the functions of CRISPR spacer acquisition are highlight in red dot.

Extended Data Fig. S9: Sequences alignment of Cas2 protein. Conserved amino acid residues involved in Mg^{2+} binding site is highlight in red dot.

Extended Data Fig. S10: Structure alignment of predicted anti-CRISPR protein (Acr) of Chiyouvirus with reference AcrIF24. Models of Bathyarchaeia virus encoding Acr is colored in green, AcrIF24 NTD is colored in wheat. AcrIF24 CTD is colored in palegcyan.

Table S1. Detailed statistics of Bathyarchaeia MAGs.

Table S2. Viral contigs targeted by Bathyarchaeia spacers.

Table S3. Classification of Bathyarchaeia viruses.

Table S4. Annotations of proteins encoded by Bathyarchaeia viruses by DRAM-v

554 **and HHblits.**

555

556 **Table S5. Comparison of Bathyarchaeia virus encoding protein with other archaea**

557 **viruses.**

558

559 **References**

560

561 1. Inagaki F*, et al.* Biogeographical distribution and diversity of microbes in methane
562 hydrate-bearing deep marine sediments on the Pacific Ocean Margin. *Proc. Natl Acad.*
563 *Sci. USA* **103**, 2815-2820 (2006).
564
565 2. Kubo K, Lloyd KG, F Biddle J, Amann R, Teske A, Knittel K. Archaea of the
566 Miscellaneous Crenarchaeotal Group are abundant, diverse and widespread in marine
567 sediments. *ISME J.* **6**, 1949-1965 (2012).
568
569 3. Lloyd KG*, et al.* Predominant archaea in marine sediments degrade detrital proteins.
570 *Nature* **496**, 215-218 (2013).
571
572 4. Lazar CS, Biddle JF, Meador TB, Blair N, Hinrichs KU, Teske AP. Environmental
573 controls on intragroup diversity of the uncultured benthic archaea of the miscellaneous
574 C renarchaeotal group lineage naturally enriched in anoxic sediments of the W hite O
575 ak R iver estuary (N orth C arolina, USA). *Environ. Microbiol.* **17**, 2228-2238 (2015).
576
577 5. Zhou Z, Pan J, Wang F, Gu JD, Li M. Bathyarchaeota: globally distributed metabolic
578 generalists in anoxic environments. *FEMS Microbiol. Rev.* **42**, 639-655 (2018).
579
580 6. He Y*, et al.* Genomic and enzymatic evidence for acetogenesis among multiple
581 lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments. *Nat.*
582 *Microbiol.* **1**, 16035 (2016).
583
584 7. Evans PN*, et al.* Methane metabolism in the archaeal phylum Bathyarchaeota revealed
585 by genome-centric metagenomics. *Science* **350**, 434-438 (2015).
586
587 8. Pan J*, et al.* Genomic and transcriptomic evidence of light-sensing, porphyrin
588 biosynthesis, Calvin-Benson-Bassham cycle, and urea production in Bathyarchaeota.
589 *Microbiome* **8**, 43 (2020).
590

591 9. Qi Y-L, *et al.* Comparative Genomics Reveals Thermal Adaptation and a High Metabolic
592 Diversity in “Candidatus Bathyarchaeia”. *mSystems* **6**, e00252-00221 (2021).
593
594 10. Yu T, Wu W, Liang W, Lever MA, Hinrichs K-U, Wang F. Growth of sedimentary
595 Bathyarchaeota on lignin as an energy source. *Proc. Natl Acad. Sci. USA* **115**, 6022
596 (2018).
597
598 11. Zhou Z, *et al.* Genomic and transcriptomic insights into the ecology and metabolism of
599 benthic archaeal cosmopolitan, Thermoprofundales (MBG-D archaea). *ISME J.* **13**,
600 885-901 (2019).
601
602 12. Chevallereau A, Pons BJ, van Houte S, Westra ER. Interactions between bacterial and
603 phage communities in natural environments. *Nat. Rev. Microbiol.* **20**, 49-62 (2022).
604
605 13. López-García P, *et al.* Metagenome-derived virus-microbe ratios across ecosystems.
606 *ISME J.*, (2023).
607
608 14. Jiao N, *et al.* Microbial production of recalcitrant dissolved organic matter: long-term
609 carbon storage in the global ocean. *Nat. Rev. Microbiol.* **8**, 593-599 (2010).
610
611 15. Brussaard CPD, *et al.* Global-scale processes with a nanoscale drive: the role of
612 marine viruses. *ISME J.* **2**, 575-578 (2008).
613
614 16. Fang X, *et al.* Transcriptomic responses of the marine cyanobacterium
615 Prochlorococcus to viral lysis products. *Environ. Microbiol.* **21**, 2015-2028 (2019).
616
617 17. Xiao X, *et al.* Viral Lysis Alters the Optical Properties and Biological Availability of
618 Dissolved Organic Matter Derived from Prochlorococcus Picocyanobacteria. *Appl.*
619 *Environ. Microbiol.* **87**, (2021).
620
621 18. Mann NH, Cook A, Millard A, Bailey S, Clokie M. Bacterial photosynthesis genes in a
622 virus. *Nature* **424**, 741-741 (2003).
623
624 19. Millard A, Clokie MRJ, Shub DA, Mann NH. Genetic organization of the psbAD region
625 in phages infecting marine Synechococcus strains. *Proc. Natl Acad. Sci. USA* **101**,
626 11007-11012 (2004).
627
628 20. Anantharaman K, Duhaime MB, Breier JA, Wendt KA, Toner BM, Dick GJ. Sulfur
629 oxidation genes in diverse deep-sea viruses. *Science* **344**, 757-760 (2014).
630
631 21. Koonin EV, Makarova KS. CRISPR-Cas: an adaptive immunity system in prokaryotes.
632 *F1000 Biol Rep* **1**, 95 (2009).
633
634 22. Barrangou R, *et al.* CRISPR provides acquired resistance against viruses in

635 prokaryotes. *Science* **315**, 1709-1712 (2007).
636
637 23. Dion MB, Plante P-L, Zufferey E, Shah SA, Corbeil J, Moineau S. Streamlining CRISPR
638 spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids*
639 *Res.* **49**, 3127-3138 (2021).
640
641 24. Pawluk A, Davidson AR, Maxwell KL. Anti-CRISPR: discovery, mechanism and function.
642 *Nat. Rev. Microbiol.* **16**, 12-17 (2018).
643
644 25. Mendoza SD*, et al.* A bacteriophage nucleus-like compartment shields DNA from
645 CRISPR nucleases. *Nature* **577**, 244-248 (2020).
646
647 26. Malone LM*, et al.* A jumbo phage that forms a nucleus-like structure evades CRISPR–
648 Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nat. Microbiol.* **5**,
649 48-55 (2020).
650
651 27. Borges AL, Davidson AR, Bondy-Denomy J. The Discovery, Mechanisms, and
652 Evolutionary Impact of Anti-CRISPRs. *Annu Rev Virol* **4**, 37-59 (2017).
653
654 28. Marino ND, Pinilla-Redondo R, Csörgő B, Bondy-Denomy J. Anti-CRISPR protein
655 applications: natural brakes for CRISPR-Cas technologies. *Nat. Methods* **17**, 471-479
656 (2020).
657
658 29. Krupovic M, Cvirkaite-Krupovic V, Iranzo J, Prangishvili D, Koonin EV. Viruses of
659 archaea: Structural, functional, environmental and evolutionary genomics. *Virus Res*
660 **244**, 181-193 (2018).
661
662 30. Medvedeva S*, et al.* Three families of Asgard archaeal viruses identified in
663 metagenome-assembled genomes. *Nat. Microbiol.* **7**, 962-973 (2022).
664
665 31. Rambo IM, Langwig MV, Leão P, De Anda V, Baker BJ. Genomes of six viruses that
666 infect Asgard archaea from deep-sea sediments. *Nat. Microbiol.* **7**, 953-961 (2022).
667
668 32. Tamarit D*, et al.* A closed Candidatus Odinarchaeum chromosome exposes Asgard
669 archaeal viruses. *Nat. Microbiol.* **7**, 948-952 (2022).
670
671 33. Laso-Pérez R*, et al.* Evolutionary diversification of methanotrophic ANME-1 archaea
672 and their expansive virome. *Nat Microbiol* **8**, 231-245 (2023).
673
674 34. Zhou Y, Zhou L, Yan S, Chen L, Krupovic M, Wang Y. Diverse viruses of marine archaea
675 discovered using metagenomics. *Environ. Microbiol.* **25**, 367-382 (2023).
676
677 35. López-Pérez M, Haro-Moreno JM, de la Torre JR, Rodriguez-Valera F. Novel
678 Caudovirales associated with Marine Group I Thaumarchaeota assembled from

679 metagenomes. *Environ. Microbiol.* **21**, 1980-1988 (2019).
680
681 36. Kim J-G*, et al.* Spindle-shaped viruses infect marine ammonia-oxidizing thaumarchaea.
682 *Proc. Natl Acad. Sci. USA* **116**, 15645-15650 (2019).
683
684 37. Ahlgren NA, Fuchsman CA, Rocap G, Fuhrman JA. Discovery of several novel,
685 widespread, and ecologically distinct marine Thaumarchaeota viruses that encode
686 amoC nitrification genes. *ISME J.* **13**, 618-631 (2019).
687
688 38. Philosof A, Yutin N, Flores-Uribe J, Sharon I, Koonin EV, Béjà O. Novel Abundant
689 Oceanic Viruses of Uncultured Marine Group II Euryarchaeota. *Curr. Biol.* **27**, 1362-
690 1368 (2017).
691
692 39. Yin X*, et al.* Subgroup level differences of physiological activities in marine
693 Lokiarchaeota. *ISME J.* **15**, 848-861 (2021).
694
695 40. Hou J*, et al.* Taxonomic and carbon metabolic diversification of Bathyarchaeia during
696 its coevolution history with early Earth surface environment. *Sci. Adv.* **9**, eadf5069
697 (2023).
698
699 41. Roux S*, et al.* IMG/VR v3: an integrated ecological and evolutionary framework for
700 interrogating genomes of uncultivated viruses. *Nucleic Acids Res.* **49**, D764-D775
701 (2021).
702
703 42. Koonin EV, Dolja VV, Krupovic M, Kuhn JH. Viruses Defined by the Position of the
704 Virosphere within the Replicator Space. *Microbiol Mol Biol Rev* **85**, e0019320 (2021).
705
706 43. Krupovic M*, et al.* Adnaviria: a New Realm for Archaeal Filamentous Viruses with Linear
707 A-Form Double-Stranded DNA Genomes. *J. Virol.* **95**, e0067321 (2021).
708
709 44. Liu Y*, et al.* Diversity, taxonomy, and evolution of archaeal viruses of the class
710 Caudoviricetes. *PLoS Biol.* **19**, e3001442 (2021).
711
712 45. Sheppard C, Werner F. Structure and mechanisms of viral transcription factors in
713 archaea. *Extremophiles* **21**, 829-838 (2017).
714
715 46. Kazlauskas D, Krupovic M, Venclovas Č. The logic of DNA replication in double-
716 stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids*
717 *Res.* **44**, 4551-4564 (2016).
718
719 47. Brosius J, Arfsten U. Primary structure of protein L19 from the large subunit of
720 Escherichia coli ribosomes. *Biochemistry* **17**, 508-516 (1978).
721
722 48. Mizuno CM*, et al.* Numerous cultivated and uncultivated viruses encode ribosomal

723 proteins. *Nat Commun* **10**, 752 (2019).
724
725 49. Wang F*, et al.* Spindle-shaped archaeal viruses evolved from rod-shaped ancestors to
726 package a larger genome. *Cell* **185**, 1297-1307.e1211 (2022).
727
728 50. Quemin Emmanuelle RJ*, et al.* Sulfolobus Spindle-Shaped Virus 1 Contains
729 Glycosylated Capsid Proteins, a Cellular Chromatin Protein, and Host-Derived Lipids.
730 *J. Virol.* **89**, 11681-11691 (2015).
731
732 51. Zhan Z, Zhou J, Huang L. Site-Specific Recombination by SSV2 Integrase: Substrate
733 Requirement and Domain Functions. *J. Virol.* **89**, 10934-10944 (2015).
734
735 52. Albers S-V, Meyer BH. The archaeal cell envelope. *Nat. Rev. Microbiol.* **9**, 414-426
736 (2011).
737
738 53. Khomyakova MA, Merkel AY, Mamiy DD, Klyukina AA, Slobodkin AI. Phenotypic and
739 genomic characterization of Bathyarchaeum tardum gen. nov., sp. nov., a cultivated
740 representative of the archaeal class Bathyarchaeia. *Front. Microbiol.* **14**, (2023).
741
742 54. Baquero DP*, et al.* New virus isolates from Italian hydrothermal environments
743 underscore the biogeographic pattern in archaeal virus communities. *ISME J.* **14**, 1821-
744 1833 (2020).
745
746 55. Wang F*, et al.* Structures of filamentous viruses infecting hyperthermophilic archaea
747 explain DNA stabilization in extreme environments. *Proc. Natl Acad. Sci. USA* **117**,
748 19643-19652 (2020).
749
750 56. Quemin ER*, et al.* First insights into the entry process of hyperthermophilic archaeal
751 viruses. *J. Virol.* **87**, 13379-13385 (2013).
752
753 57. Prangishvili D, Koonin EV, Krupovic M. Genomics and biology of Rudiviruses, a model
754 for the study of virus-host interactions in Archaea. *Biochem. Soc. Trans.* **41**, 443-450
755 (2013).
756
757 58. Baquero DP, Liu Y, Wang F, Egelman EH, Prangishvili D, Krupovic M. Structure and
758 assembly of archaeal viruses. *Adv Virus Res* **108**, 127-164 (2020).
759
760 59. Pinilla-Redondo R*, et al.* Type IV CRISPR–Cas systems are highly diverse and
761 involved in competition between plasmids. *Nucleic Acids Res.* **48**, 2000-2012 (2020).
762
763 60. Faure G*, et al.* CRISPR–Cas in mobile genetic elements: counter-defence and beyond.
764 *Nat. Rev. Microbiol.* **17**, 513-525 (2019).
765
766 61. Moya-Beltrán A*, et al.* Evolution of Type IV CRISPR-Cas Systems: Insights from

767 CRISPR Loci in Integrative Conjugative Elements of Acidithiobacillia. *Crispr J.* **4**, 656-
768 672 (2021).
769
770 62. Al-Shayeb B*, et al.* Diverse virus-encoded CRISPR-Cas systems include streamlined
771 genome editors. *Cell* **185**, 4574-4586.e4516 (2022).
772
773 63. Zhou Y*, et al.* Structure of a type IV CRISPR-Cas ribonucleoprotein complex. *iScience*
774 **24**, 102201 (2021).
775
776 64. Medvedeva S, Liu Y, Koonin EV, Severinov K, Prangishvili D, Krupovic M. Virus-borne
777 mini-CRISPR arrays are involved in interviral conflicts. *Nat. Commun.* **10**, 5204 (2019).
778
779 65. Shiimori M, Garrett SC, Graveley BR, Terns MP. Cas4 Nucleases Define the PAM,
780 Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol. Cell* **70**,
781 814-824.e816 (2018).
782
783 66. Hudaiberdiev S, Shmakov S, Wolf YI, Terns MP, Makarova KS, Koonin EV.
784 Phylogenomics of Cas4 family nucleases. *BMC Evol. Biol.* **17**, 232 (2017).
785
786 67. Zhang Z, Pan S, Liu T, Li Y, Peng N. Cas4 Nucleases Can Effect Specific Integration
787 of CRISPR Spacers. *J. Bacteriol.* **201**, (2019).
788
789 68. Guo Y, Kragelund BB, White MF, Peng X. Functional characterization of a conserved
790 archaeal viral operon revealing single-stranded DNA binding, annealing and nuclease
791 activities. *J Mol Biol* **427**, 2179-2191 (2015).
792
793 69. Wandera KG*, et al.* Anti-CRISPR prediction using deep learning reveals an inhibitor of
794 Cas13b nucleases. *Molecular cell* **82**, 2714-2726. e2714 (2022).
795
796 70. Yang L*, et al.* Insights into the inhibition of type I-F CRISPR-Cas system by a
797 multifunctional anti-CRISPR protein AcrIF24. *Nat. Commun.* **13**, 1931 (2022).
798
799 71. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB:
800 an ongoing census of bacterial and archaeal diversity through a phylogenetically
801 consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids*
802 *Res.* **50**, D785-D794 (2022).
803
804 72. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly
805 classification with the genome taxonomy database. *Bioinform.* **38**, 5315-5316 (2022).
806
807 73. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein
808 sequence and structural alignment. *Nucleic Acids Res.* **47**, W5-W10 (2019).
809
810 74. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated

811 alignment trimming in large-scale phylogenetic analyses. *Bioinform.* **25**, 1972-1973
812 (2009).
813
814 75. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing
815 the quality of microbial genomes recovered from isolates, single cells, and
816 metagenomes. *Genome Res.* **25**, 1043-1055 (2015).
817
818 76. Parks DH*, et al.* A standardized bacterial taxonomy based on genome phylogeny
819 substantially revises the tree of life. *Nat Biotechnol* **36**, 996-1004 (2018).
820
821 77. Tesson F*, et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes.
822 *Nat. Commun.* **13**, 2561 (2022).
823
824 78. Russel J, Pinilla-Redondo R, Mayo-Muñoz D, Shah SA, Sørensen SJ.
825 CRISPRCasTyper: Automated Identification, Annotation, and Classification of
826 CRISPR-Cas Loci. *CRISPR J.* **3**, 462-469 (2020).
827
828 79. Bland C*, et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of
829 clustered regularly interspaced palindromic repeats. *BMC Bioinform.* **8**, 209 (2007).
830
831 80. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND.
832 *Nat. Methods* **12**, 59-60 (2015).
833
834 81. Camacho C*, et al.* BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421
835 (2009).
836
837 82. Guo J*, et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse
838 DNA and RNA viruses. *Microbiome* **9**, 37 (2021).
839
840 83. Nayfach S, Camargo AP, Schulz F, Eloe-Fadrosh E, Roux S, Kyrpides NC. CheckV
841 assesses the quality and completeness of metagenome-assembled viral genomes. *Nat.*
842 *Biotechnol.* **39**, 578-585 (2021).
843
844 84. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinform.* **30**, 2068-2069
845 (2014).
846
847 85. Shaffer M*, et al.* DRAM for distilling microbial metabolism to automate the curation of
848 microbiome function. *Nucleic Acids Res.* **48**, 8883-8900 (2020).
849
850 86. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3
851 for fast remote homology detection and deep protein annotation. *BMC Bioinform.* **20**,
852 473 (2019).
853
854 87. Grant JR*, et al.* Proksee: in-depth characterization and visualization of bacterial

genomes. *Nucleic Acids Res.* **51**, W484-W492 (2023).

88. Bin Jang H*, et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632-639 (2019).

89. Shannon P*, et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504 (2003).

90. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. *Bioinform.* **33**, 2379-2380 (2017).

91. Meier-Kolthoff JP, Göker M. VICTOR: genome-based phylogeny and classification of prokaryotic viruses. *Bioinform.* **33**, 3396-3404 (2017).

92. Jumper J*, et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

93. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679-682 (2022).

94. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).

95. Makarova KS*, et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67-83 (2020).

96. Hatcher EL*, et al.* Virus Variation Resource - improved response to emergent viral outbreaks. *Nucleic Acids Res.* **45**, D482-d490 (2017).

97. Xie J, Chen Y, Cai G, Cai R, Hu Z, Wang H. Tree Visualization By One Table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. *Nucleic Acids Res.* **51**, W587-W592 (2023).

98. Schmartz GP*, et al.* PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res.* **50**, D273-d278 (2022).

99. Pourcel C*, 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.* **48**, D535-D544 (2020).

100. Bondy-Denomy J*, et al.* A Unified Resource for Tracking Anti-CRISPR Names. *Crispr J.* **1**, 304-305 (2018).

101. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the

TM-score. *Nucleic Acids Res.* **33**, 2302-2309 (2005).

102. Ren M*, et al.* Phylogenomics suggests oxygen availability as a driving force in Thaumarchaeota evolution. *ISME J.* **13**, 2150-2161 (2019).

103. Dobson L, Reményi I, Tusnády GE. CCTOP: a Consensus Constrained TOPology prediction web server. *Nucleic Acids Res.* **43**, W408-W412 (2015).

104. Kim D, Park S, Chun J. Introducing EzAAI: a pipeline for high throughput calculations of prokaryotic average amino acid identity. *Journal of Microbiology* **59**, 476-480 (2021).

Fig. 1

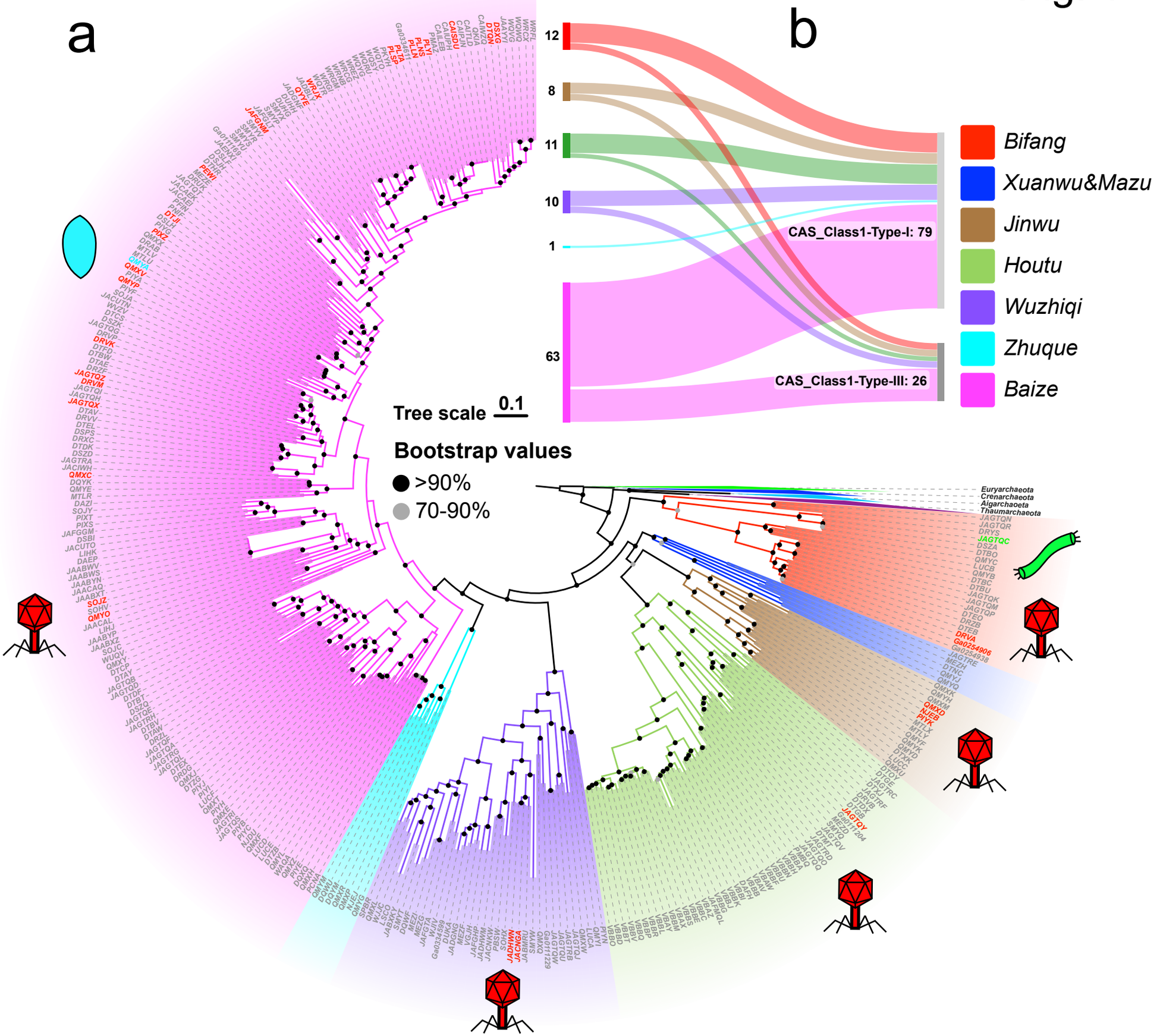

a
b
12
8
11
10
1
63
CAS_Class1-Type-I: 79
CAS_Class1-Type-III: 26
Bifang
Xuanwu&Mazu
Jinwu
Houtu
Wuzhiqi
Zhuque
Baize
Tree scale 0.1
Bootstrap values
>90%
70-90%
Euryarchaeota
Crenarchaeota
Aigarchaeota
Thaumarchaeota

Fig. 2

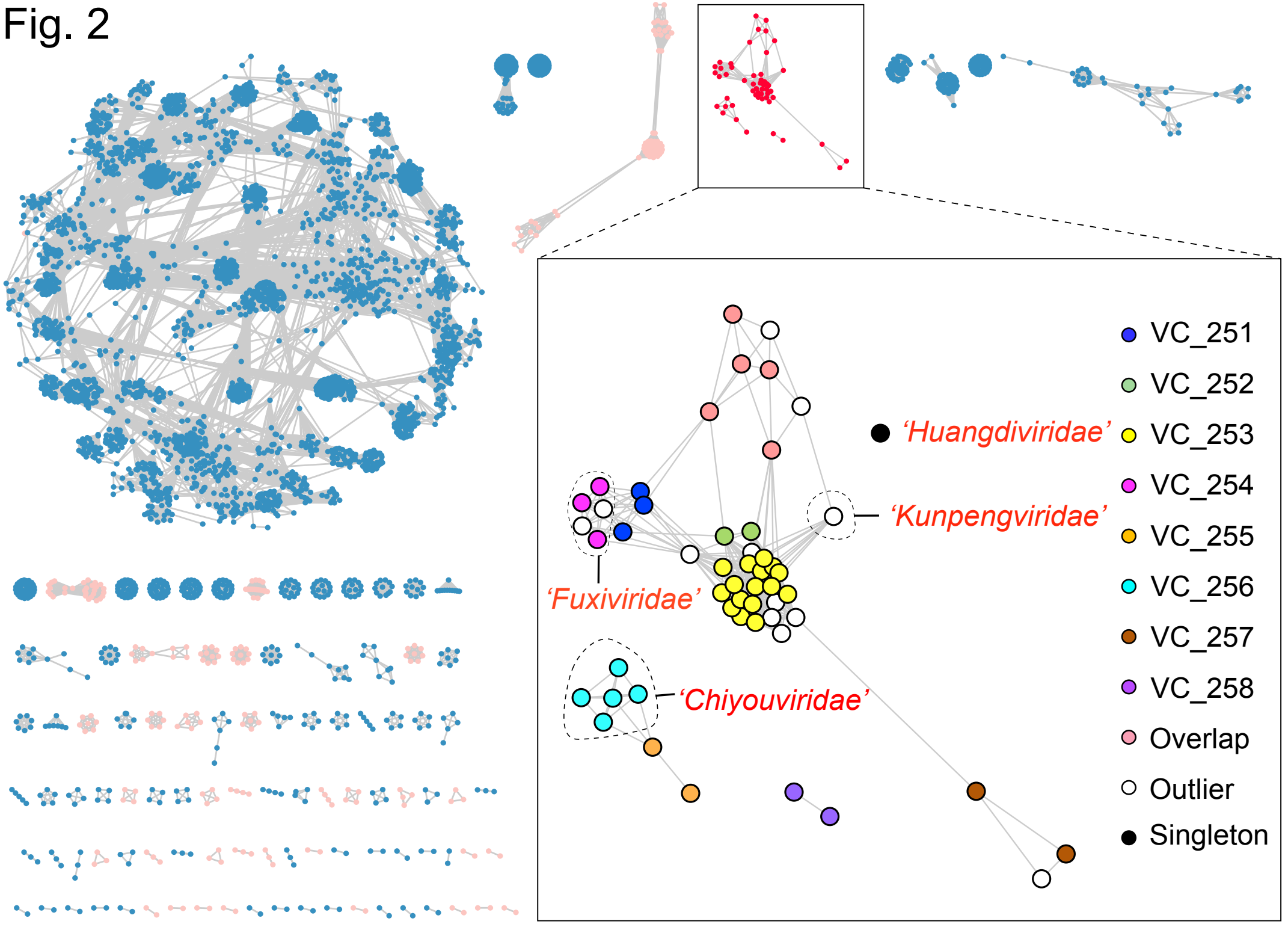
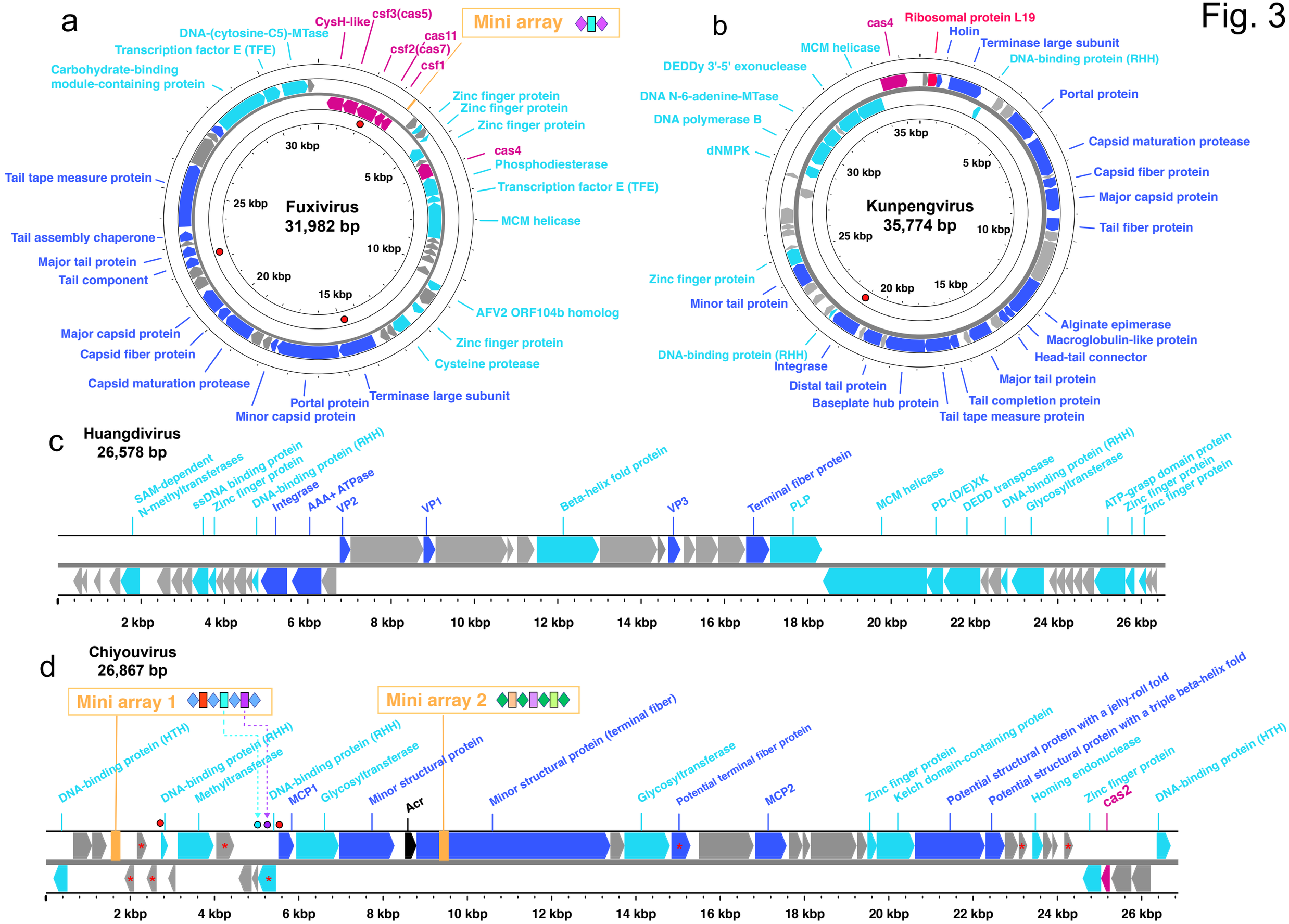

'Huangdiviridae'
'Kunpengviridae'
'Fuxiviridae'
'Chiyouviridae'
VC_251
VC_252
VC_253
VC_254
VC_255
VC_256
VC_257
VC_258
Overlap
Outlier
Singleton

Fig. 3


a
Fuxivirus
31,982 bp
Mini array
CysH-like
csf3(cas5)
cas11
csf2(cas7)
csf1
DNA-(cytosine-C5)-MTase
Transcription factor E (TFE)
Carbohydrate-binding module-containing protein
Zinc finger protein
Zinc finger protein
Zinc finger protein
cas4
Phosphodiesterase
Transcription factor E (TFE)
MCM helicase
AFV2 ORF104b homolog
Zinc finger protein
Cysteine protease
Terminase large subunit
Portal protein
Minor capsid protein
Capsid maturation protease
Capsid fiber protein
Major capsid protein
Tail component
Major tail protein
Tail assembly chaperone
Tail tape measure protein
30 kbp
5 kbp
25 kbp
10 kbp
20 kbp
15 kbp
b
Kunpengvirus
35,774 bp
cas4
Ribosomal protein L19
Holin
Terminase large subunit
DNA-binding protein (RHH)
MCM helicase
DEDDy 3'-5' exonuclease
DNA N-6-adenine-MTase
DNA polymerase B
dNMPK
Portal protein
Capsid maturation protease
Capsid fiber protein
Major capsid protein
Tail fiber protein
Zinc finger protein
Minor tail protein
DNA-binding protein (RHH)
Integrase
Distal tail protein
Baseplate hub protein
Tail tape measure protein
Tail completion protein
Major tail protein
Head-tail connector
Macroglobulin-like protein
Alginate epimerase
35 kbp
5 kbp
30 kbp
10 kbp
25 kbp
20 kbp
15 kbp
c
Huangdivirus
26,578 bp
SAM-dependent N-methyltransferases
ssDNA binding protein
Zinc finger protein
DNA-binding protein (RHH)
Integrase
AAA+ ATPase
VP2
VP1
Beta-helix fold protein
VP3
Terminal fiber protein
PLP
MCM helicase
PD-(D/E)XK
DEDD transposase
DNA-binding protein (RHH)
Glycosyltransferase
ATP-grasp domain protein
Zinc finger protein
Zinc finger protein
2 kbp
4 kbp
6 kbp
8 kbp
10 kbp
12 kbp
14 kbp
16 kbp
18 kbp
20 kbp
22 kbp
24 kbp
26 kbp
d
Chiyouvirus
26,867 bp
Mini array 1
Mini array 2
DNA-binding protein (HTH)
DNA-binding protein (RHH)
Methyltransferase
DNA-binding protein (RHH)
MCP1
Glycosyltransferase
Minor structural protein
Acr
Minor structural protein (terminal fiber)
Glycosyltransferase
Potential terminal fiber protein
MCP2
Zinc finger protein
Kelch domain-containing protein
Potential structural protein with a jelly-roll fold
Potential structural protein with a triple beta-helix fold
Homing endonuclease
Zinc finger protein
cas2
DNA-binding protein (HTH)
2 kbp
4 kbp
6 kbp
8 kbp
10 kbp
12 kbp
14 kbp
16 kbp
18 kbp
20 kbp
22 kbp
24 kbp
26 kbp