Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsid-less mobile elements

Jaime Iranzo, Eugene V. Koonin, David Prangishvili, Mart Krupovic

JVI Accepted Manuscript Posted Online 28 September 2016
J. Virol. doi:10.1128/JVI.01622-16

Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsid-less mobile elements

Jaime Iranzo[a], Eugene V. Koonin[a], David Prangishvili[b], Mart Krupovic[b,#]

a – National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894

b – Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, 25 rue du Docteur Roux, 75015 Paris, France

Corresponding author: krupovic@pasteur.fr

Running title: Bipartite network analysis of the archaeal virosphere

Word count

Abstract: 245

Importance: 107

Text: 8116

Abstract

Archaea and particularly hyperthermophilic crenarchaea are hosts to many unusual viruses with diverse virion shapes and distinct gene compositions. As is typical of viruses in general, there are no universal genes in the archaeal virosphere. Therefore, to obtain a comprehensive picture of the evolutionary relationships between viruses, network analysis methods are more productive than traditional phylogenetic approaches. Here we present a comprehensive comparative analysis of genomes and proteomes from all currently known taxonomically classified and unclassified, cultivated and uncultivated archaeal viruses. We constructed a bipartite network of archaeal viruses that includes two classes of nodes, the genomes and gene families that connect them. Dissection of this network using formal community detection methods reveals strong modularity with 10 distinct modules and 3 putative supermodules. However, compared to the previously analyzed similar networks of eukaryotic and bacterial viruses, the archaeal virus network is sparsely connected. With the exception of the tailed viruses related to the bacteriophages of the order *Caudovirales* and the families *Turriviridae* and *Sphaerolipoviridae* that are linked to a distinct supermodule of eukaryotic viruses, there are few connector genes shared by different archaeal virus modules. In contrast, most of these modules include, in addition to viruses, capsid-less mobile elements, emphasizing tight evolutionary connections between the two types of entities in archaea. The relative contributions of distinct evolutionary origins, in particular from non-viral elements, and insufficient sampling to the sparsity of the archaeal virus network remain to be determined by further exploration of the archaeal virosphere.

Importance

Viruses infecting archaea are among the most mysterious denizens of the virosphere. Many of these viruses display no genetic or even morphological relationship to viruses of bacteria and eukaryotes, raising

questions regarding their origins and position in the global virosphere. Analysis of 5740 protein sequences from 116 genomes allowed dissection of the archaeal virus network and showed that most groups of the archaeal viruses are evolutionarily connected to capsid-less mobile genetic elements, including various plasmids and transposons. This finding could reflect actual independent origins of the distinct groups of archaeal viruses from different non-viral elements, providing important insights into the emergence and evolution of the archaeal virome.

Journal of Virology JVI Journal of Virology

49 **Introduction**

50 Viruses infecting archaea are among the most mysterious denizens of the virosphere. Archaeal viruses
51 display a rich diversity of virion morphotypes and can be broadly divided into two categories: those that
52 are structurally and genetically related to bacterial or eukaryotic viruses and those that are archaea-specific
53 (1-4). The cosmopolitan fraction of archaeal viruses includes (i) head-tailed viruses of the order
54 *Caudovirales*, with the 3 included families: *Siphoviridae, Myoviridae* and *Podoviridae*; (ii) tailless icosahedral
55 viruses of the families *Sphaerolipoviridae* and *Turriviridae*, (iii) enveloped pleomorphic viruses of the family
56 *Pleolipoviridae*. All of these viruses, except for those of the family *Turriviridae*, propagate in members of
57 the archaeal phylum Euryarchaeota, whereas most of the archaea-specific virus groups infect
58 hyperthermophilic organisms of the phylum Crenarchaeota.

59 The order *Caudovirales* contains three families, *Siphoviridae, Myoviridae* and *Podoviridae*, and includes
60 both bacterial and archaeal viruses. Members of the *Caudovirales* feature icosahedral capsids and helical
61 tails attached to one of the vertices of the capsid. These viruses have been largely isolated from
62 hyperhalophilic archaea (order Halobacteriales) although one tailed archaeal virus has been isolated from a
63 methanogen host (order Methanobacteriales) (5-7). However, related proviruses have been characterized
64 from a broader range of archaea which, in addition to members of the Halobacteriales and
65 Methanobacteriales, includes euryarchaea from the orders Methanococcales and Methanosarcinales as
66 well as the phylum Thaumarchaeota (8-10).

67 The recently created family *Sphaerolipoviridae* includes viruses with tail-less icosahedral capsids and
68 internal membranes. Members of the genera *Alpha-* and *Betasphaerolipovirus* infect halophilic archaea,
69 whereas those of the genus *Gammasphaerolipovirus* propagate in bacteria of the genus *Thermus* (11).
70 Aphasphaerolipoviruses possess linear dsDNA genomes (12, 13), whereas betasphaerolipoviruses and
71 gammasphaerolipoviruses encapsidate circular genomes (14, 15). Nevertheless, virion organization and

72 morphogenesis are similar among viruses in all three genera; all these viruses encode two major capsid
73 proteins (MCP) with the single jelly-roll fold and apparently encapsidate their genomes using homologous
74 A32-like packaging ATPases (16, 17).

75 Viruses of the *Turriviridae* family resemble sphaerolipoviruses in the overall virion organization but instead
76 of the two capsid proteins employ one MCP with the double jelly-roll fold and one minor capsid protein
77 with the single jelly-roll fold (18). Similar to sphaerolipoviruses, turriviruses encode A32-like genome
78 packaging ATPases. Structurally similar viruses infect hosts in all three domains of life, suggesting a long
79 evolutionary history of this supergroup of viruses (19, 20). Turriviruses are known to infect
80 hyperthermophilic crenarchaea of the order Sulfolobales but proviruses encoding homologous MCPs and
81 genome packaging ATPases have been described also in organisms from other orders of the Crenarchaota
82 and the phylum Euryarchaeota (21-23). Pleomorphic viruses of the family *Pleolipoviridae* are unique in that
83 genetically closely related members encapsidate either single-stranded (ss) or dsDNA genomes (24, 25).
84 Viruses with morphologically similar virions (with both ssDNA and dsDNA genomes) have been described in
85 bacteria (members of the family *Plasmaviridae* and some unclassified phages) although the exact
86 evolutionary relationship between these bacterial and archaeal viruses remains unclear.

87 Archaea-specific viruses are classified into 10 families (2). Arguably, the most unexpected among these are
88 members of the family *Ampullaviridae* with bottle-shaped virions. The family is currently represented by a
89 single isolate, Acidianus bottle-shaped virus (ABV) (26), but two additional complete ABV-like genomes
90 have been recently assembled from metagenomic data (27). Ampullaviruses contain linear dsDNA genomes
91 with terminal inverted repeats which appear to be replicated by the virus-encoded protein-primed DNA
92 polymerases of the B family (26). Crenarchaea are infected by a range of filamentous viruses which can be
93 flexible or rigid, long or short, contain dsDNA or ssDNA genomes. These viruses are classified into 5
94 families: *Rudiviridae, Lipothrixviridae, Tristromaviridae, Clavaviridae* and *Spiraviridae*. Rod-shaped, non-
95 enveloped rudiviruses and flexible, enveloped lipothrixviruses share a considerable fraction of genes,

including those encoding the major capsid proteins. In recognition of this evolutionary relationship, the two families are unified within the order *Ligamenvirales* (28). Another family of enveloped filamentous viruses is *Tristromaviridae*; these viruses do not share genes with other archaeal viruses and have a virion organization which is more complex than that of lipothrixviruses (29). The family *Clavaviridae* includes a single virus isolate with bacilliform virions. Aeropyrum pernix bacilliform virus 1 (APBV1) contains a circular dsDNA genome of 5 kb and is among the smallest dsDNA viruses known (30). By contrast, the spiravirus Acidianus coil-shaped virus (ACV) contains by far the largest (~25 kb) genome among known ssDNA viruses. The ACV virion is organized as a coil which is prone to contraction and stiffening (31).

Among the most widespread archaeal viruses are those with spindle-shaped virions. Such viruses are thus far exclusive to archaea and have been detected in diverse habitats, including deep sea hydrothermal vents, hypersaline environments, anoxic freshwaters, cold Antarctic lakes, terrestrial hot springs, and acidic mines (32). There are two lineages of spindle-shaped viruses which appear to be evolutionarily unrelated. The first group includes crenarchaeal viruses of the family *Fuselloviridae*, exemplified by the Sulfolobus spindle-shaped virus 1 (SSV1), as well as several unclassified viruses infecting crenarchaeal and euryarchaeal hosts (33-38). All these viruses contain relatively small dsDNA genomes (<20 kb) and share specific MCPs. Thus, it has been suggested that crenarchaeal and euryarchaeal spindle-shaped viruses could be unified into one family (32). The other group includes considerably larger spindle-shaped viruses which, unlike the SSV1-like viruses, are decorated with one or two long, tail-like appendages. The latter can develop either intracellularly or in the extracellular medium. These viruses possess the largest dsDNA genomes among crenarchaeal viruses (up to 76 kb). Acidianus two-tailed virus (ATV) is currently the only classified representative of this virus group and the type species of the family *Bicaudaviridae* (39). Many morphologically similar viruses have been isolated and related genomes have been assembled from metagenomic data (27, 40-44). Unlike ATV, these viruses typically contain one tail and are often referred to as “monocaudaviruses”; however, such a taxon currently has no official standing, and accordingly, these

viruses remain unclassified. Finally, two additional families, *Guttaviridae* and *Globuloviridae*, include viruses with dsDNA genomes and droplet-shaped or spherical virions, respectively (2).

The vast majority of archaeal viruses contain dsDNA genomes, whereas viruses with ssDNA genomes are rare and those with RNA genomes have not been isolated although tentative indications of the possible existence of RNA viruses infecting hyperthermophilic crenarchaea have been obtained from metagenomic data (45). Archaeal viruses typically contain a large fraction of genes of unknown function. This is especially true for creanarchaeal viruses. A global comparative genomic analysis of archaeal viruses, performed a decade ago, has revealed a small pool of genes shared by overlapping subsets of archaeal viruses as well as several genes with prokaryotic homologs (3). Furthermore, a growing body of data indicates that archaeal viruses often share genes with non-viral selfish replicons, such as plasmids and transposons. During the past few years, a number of new archaeal viruses were isolated and several complete new genomes of uncultivated archaeal viruses were obtained. This prompted us to systematically reevaluate the relationships between all known groups of archaeal viruses using bipartite network analysis. The results of this analysis substantially extend the understanding of the evolution of the archaeal virosphere and emphasize the important contribution of non-viral elements in this process.

Methods

Sequences

Protein sequences were collected from the NCBI Genome database for all available genomes of archaeal viruses. Specifically, we collected genomes of viruses belonging to the families *Ampullaviridae*, *Bicaudaviridae*, *Clavaviridae*, *Fuselloviridae*, *Globuloviridae*, *Guttaviridae*, *Lipothrixviridae*, *Pleolipoviridae*,

142 *Rudiviridae*, *Sphaerolipoviridae* (note that viruses from genera *Alfa*- and *Betasphaerolipovirus* infect
143 archaea, whereas those from the genus *Gammasphaerolipovirus* infect bacteria), *Spiraviridae*,
144 *Tristromaviridae* and *Turriviridae*, as well as members of the order *Caudovirales* (families *Siphoviridae*,
145 *Podoviridae* and *Myoviridae*) that infect Archaea. This data set was complemented with sequences of
146 unclassified archaeal viruses, including those assembled from metagenomic data, as well as previously
147 described proviruses (8, 9, 21, 46), plasmids and casposons known to share genes with archaeal viruses. In
148 total, the initial data set contained 5740 protein sequences from 116 genomes.

149 *Classification of genes into homologous families*

150 Following the same methodology described in (47), all protein sequences were initially clustered at 90%
151 identity and 70% coverage using CD-HIT (48) to generate a non-redundant data set. For each sequence in
152 this set, a BLASTp search (49) with composition-based statistics (50) and filtering of low complexity regions
153 was carried out against all other sequences. An e-value cutoff equal to 0.01 (database size fixed to 2e7) was
154 used to determine valid hits. The scores for those hits were subsequently collected from a BLASTp search
155 with neither composition-based statistics nor low complexity filter. The set of scored BLAST hits defined a
156 weighted sequence similarity network, that we partitioned with Infomap (51) (100 trials, 2-level hierarchy)
157 in order to generate preliminary groups of homologous genes. In the next step, we applied profile analysis
158 to find and merge groups of related sequences. To that purpose, sequences in each group were aligned
159 with Muscle (52) (default parameters) and the alignments were used to predict secondary structure and
160 build profiles with the tools 'addss' and 'hhmake' available within the HH-suite package (53). The collection
161 of profiles was enriched with those generated in (47) for a large number of (non-archaeal) dsDNA viruses.
162 Profile-profile comparisons were carried out using HHsearch (54). To accept or reject hits, we applied the
163 same heuristics as in (47): hits with probability greater than 0.90 were accepted if they covered at least
164 50% of the length of the profile; additionally, hits with coverage 20% or greater were also accepted if their
165 probability was greater than 0.99 and their length greater than 100 aa. This pipeline rendered a total of

2931 clusters of homologous sequences of which 938 comprised multiple sequences and 1993 were singletons (ORFans).

Some groups of homologous sequences were manually curated to account for cases of remote homology that, despite being well supported by previous research, remained undetected by our automatic analysis. Such highly diverged but well supported homology occurs, for example, in capsid proteins of different groups of viruses. The main groups that had to be manually merged included capsid proteins with the HK97-like fold, caudoviral prohead maturation proteases of the U9/U35 family, capsid proteins of fuselloviruses, capsid proteins of *Ligamenvirales*, and integral membrane protein of pleolipoviruses. Henceforth, we use the term *gene family* to refer to the manually curated groups of homologous sequences.

Identification of core genes

We defined core genes as those genes that tend to be maintained in the genomes of closely related lineages in the course of evolution. According to such definition, core genes were identified by calculating the evolutionary loss rates of every gene and selecting the genes with loss rates below a given threshold. In the absence of reliable species trees, the loss rate of a gene was estimated by assuming a pure-loss evolutionary scenario, in which genomes that diverge from a common ancestor lose genes at a constant, gene-specific rate. Under this scenario, maximum likelihood estimates for the gene loss rates can be easily computed provided that (i) there is a collection of pairs of genomes, with the gene of interest present in at least one member of each pair, and (ii) the times elapsed since the last common ancestors of each pair are known. For the former, we used all possible pairs of genomes under study, excluding those pairs in which the gene of interest was absent or whose members are markedly unrelated (see below). As a proxy for the latter, and consistently with the assumption of a pure-loss evolutionary model, we computed the distance between every pair of genomes as $D_{ij} = -\ln(S_{ij}/\sqrt{N_i N_j})$, where S_{ij} is the number of families shared by

both genomes and N_i, N_j are the number of families in each genome (47). Relative to this distance, the time from the last common ancestor can be simply expressed as $t = D_{ij}/2$. Then, we used the pure-loss evolutionary model to estimate the loss rate of every gene family. According to this model, the probability that a gene family that was present in the common ancestor is still present in a single genome after a time t is $P_1 = e^{-r\,t}$, where r is the loss rate of the family relative to the average divergence rate of genomes. In the case of a pair of genomes, the probability that both members of the pair maintain the gene family conditioned on its presence in the last common ancestor is $P_{11} = e^{-r\,D_{ij}}/Z$. Similarly, the probability for the family being maintained in one genome of the pair and lost in the other is $P_{10} = 2\,e^{-rD_{ij}/2}\,(1 - e^{-rD_{ij}/2})/Z$, where $Z = P_{10} + P_{11}$ is a normalization factor. Pairs of genomes that lack any representative of the family of interest were discarded because there is no guarantee that such family was present in their common ancestor. Moreover, only those gene families with three or more appearances were considered. We used the expressions for the probabilities P_{10}, P_{11} and D_{ij} to calculate a maximum likelihood estimate of the family-specific loss rate r. The presence of one or a few shared families in otherwise unrelated genomes due to HGT could bias loss rate estimates, thus we only considered those pairs of genomes with distances $D_{ij} < 1$. Genes with loss rate $r < 1$ were assigned to the 'core'. In this way we obtained a list of 2560 core gene families, of which 180 were not classified as core in a previous analysis of the dsDNA virus world (47). Such increase in the number of detected core genes is a consequence of the deeper sampling of archaeal viral genomes, which enhances the sensitivity of the core detection algorithm. The list was completed with 12 additional core genes from the dsDNA virus world with significant presence in archaeal viruses. Supplementary Table S2 contains the list of core genes and their abundances.

Gene family *abundances* were computed based on genome-weighted contributions as previously described (55) and normalized so that an abundance equal to one implies that the family is present in all genomes of the dataset. We used the term *prevalence* to refer to the relative abundance of a gene family in a group of genomes; in computing prevalences, similarity-based genome weights were also taken into account.

Construction and analysis of the bipartite network of viruses

A bipartite network was built by connecting genome nodes to gene family nodes whenever a genome contained at least one representative of a given family. To avoid redundancy, genomes that share more than 90% of their gene content (inlcuding ORFans) were treated as a single pangenome. For practical purposes, we restricted our analysis to a reduced subset of the whole network that contained core gene families only. Moreover, three minor disconnected components encompassing, respectively, the only available genome from a member of the *Clavaviridae* (Aeropyrum pernix bacilliform virus 1), both representatives of *Tristromaviridae* (Pyrobaculum filamentous virus 1 and Thermoproteus tenax virus 1), and the globulovirus Thermoproteus tenax spherical virus 1 were excluded from further analysis.

Sets, or *modules*, of related genomes and gene families stand out by displaying a dense web of connections with members of the same module but much fewer links to genomes and gene families that do not belong to the respective module. Modularity in bipartite networks is customarily quantified by Barber's bipartite modularity index (56) which, for a given partition of the network nodes into modules, compares the observed connectivity patterns to those expected in a randomly connected network. Therefore, the modular structure of a network can be obtained by finding the partition of the network that maximizes Barber's modularity. The program Modular with default parameters (57) was used to find such optimal partition in the bipartite network consisting of genomes and core genes. Due to the stochastic nature of the module optimization algorithm, repeated runs of the algorithm on the same network typically yield different partitions with similar values of Barber's modularity. To account for this stochasticity, we ran 100 replicas of the algorithm and kept the partition with the highest modularity as the optimal partition. To evaluate the robustness of each module, we took pairs of nodes (genomes or gene families) belonging to the same module in the optimal partition and calculated the average fraction of the other 99 alternative partitions in which both nodes were grouped together. Additionally, the statistical significance of the whole partition was assessed by running 100 replicas of a null model consisting of randomly generated bipartite

networks with the same size and the same gene- and genome-degree distributions as the original network ("null model 2" provided by Modular) (58).

Provided the modular structure of the virus network, we say that a gene family is a *connector* between two modules if its prevalence in both modules is greater than exp(-1) (prevalence thresholds from 0.3 to 0.5 yield qualitatively similar results). Connector gene families were used to generate a second-order bipartite network consisting of modules and connector genes, as well as non-connector genes whose abundance exceeded the threshold in a single module. We detected supemodules by applying the module detection algorithm described above to this second-order network. As with primary modules, 100 independent replicas were carried out in order to assess the robustness of the supermodules.

The relationship between archaeal viruses belonging to the order *Caudovirales* and tailed bacteriophages was explored by connecting the archaeal virus network to the *Caudovirales* network studied in (47). To that end, we complemented the list of core genes from archaeal viruses with core genes from other dsDNA viruses, built the corresponding bipartite network of *Caudovirales* genomes and core genes, and applied the module detection algorithm to the resulting network.

Hallmark and signature genes

Hallmark genes were defined on the basis of connector genes and network supermodules. Specifically, a gene was classified as hallmark if it fulfilled two conditions: (i) to be a connector gene and (ii) to have a prevalence greater than a given threshold in at least one of the supermodules. Any prevalence threshold between 0.35 and 0.5 results in the same list of hallmark genes; thus we arbitrarily chose exp(-1) to keep consistency with the threshold used to define connector genes.

Signature genes were defined on the basis of their normalized mutual information (MI) with respect to their best and second best matching modules (47). Specifically, we required that a signature gene has MI>0.6 for the best match and MI<0.02 for the second best match.

260

261

262 **Results**

263 *The archaeal virosphere as a bipartite network of genomes and genes*

264 Predicted proteins encoded in all available genomes of archaeal viruses were classified into families of
265 homologs by sequence similarity (see Methods). The patterns of gene sharing were used to generate a
266 network of the archaeal virosphere. The network consists of two types of entities (nodes): genomes and
267 gene families. Edges connect every genome with the gene families that it contains. The result is a bipartite
268 network in which genomes are connected only through genes, and conversely, different gene families are
269 connected through genomes in which they are jointly represented. By incorporating both genes and
270 genomes, the bipartite network representation provides for a comprehensive dissection of the genomic
271 relationships among different groups of viruses.

272 To enrich the representation of certain archaeal virus families and to further investigate the evolutionary
273 connections between viruses and non-viral mobile genetic elements (MGE), we included 16 previously
274 described archaeal proviruses (related to viruses of the order *Caudovirales* and the families *Fuselloviridae*,
275 *Turriviridae and Pleolipoviridae*), 11 archaeal plasmids and 3 casposons (self-synthesizing transposons). The
276 only member of the family *Clavaviridae* (Aeropyrum pernix bacilliform virus 1) does not share genes with
277 the rest of the archaeal viruses and therefore remains separated from the network. After combining highly
278 similar genomes, the bipartite network of archaeal viruses consisted of 111 genomes and 2883 gene
279 families.

280 For efficient analysis of a complex network, it is desirable to minimize the effect of noisy connections that
281 reduce the power of most network analysis tools. In the case of a bipartite gene-genome network, such

noisy connections are generated by rare genes, low quality gene families (those containing a significant fraction of potential false hits, for example, due to short repetitive motifs) and highly mobile genes with a patchy distribution. Accordingly, we focused our analysis on a reduced version of the bipartite network that only includes "core genes", i.e. genes that tend to be retained by groups of related viruses during evolution. Throughout the rest of this work, we discuss the bipartite network composed of archaeal viral genomes and their core genes.

The bipartite network of archaeal viruses (Figure 1) includes a giant connected component that contains 107 (pro)viral genomes and 274 core gene families. Apart from that giant component and in addition to the aforementioned 'orphan' clavavirus genome, there are three genomes for which no core genes were identified, namely the two representatives of the family *Tristromaviridae* (Pyrobaculum filamentous virus 1 and Thermoproteus tenax virus 1) and the globulovirus Thermoproteus tenax spherical virus 1 (TTSV). Accordingly, these genomes remain isolated from the network. The two tristromaviruses as well as TTSV and the other globulovirus, Pyrobaculum spherical virus, share genes with each other. However, because the core detection algorithm requires that a gene be present in at least three genomes, none of those shared genes could be classified as core. Figure 1 shows that the archaeal members of the order *Caudovirales* and of the family *Sphaerolipoviridae* belong to a dense web of gene sharing with the corresponding groups of bacteriophages, whereas the other groups of archaeal viruses form well-defined clusters (*modules*) interconnected by a small number of *connector* genes. We discuss such modules and connector genes in the following sections.

Modular structure of the archaeal virus network

We applied a stochastic module detection algorithm to the archaeal virus bipartite gene-genome network. Specifically, the algorithm was run 100 times (each run was considered a replica), and the robustness of a

module was defined as the number of runs in which its members clustered together. A pronounced modular organization of the network was detected ($p<0.01$ when compared to a random network).

The network consists of 10 robust modules (Table 1 and Supplementary Table S1) most of which encompass viruses from one, or in some cases, two families (Figure 2). The size of a module in terms of the number of genomes can markedly differ from its size in terms of the number of genes. To characterize individual modules, we examined their composition with respect to both genomes and genes. In particular, *signature* genes were defined as those that are characteristic of a module, based on information theory measures (see Methods and Supplementary Table S2). Informally, signature genes are nearly exclusive to a particular module within the given network, and their relative abundance (prevalence) in such a module is close to unity. Some modules harbor numerous signature genes but others have few or none (Figure 1). We defined *connector* genes as those genes which are highly prevalent in two or more modules, effectively connecting them. The complete list of the connector genes of archaeal viruses can be found in Table 2. The modules and connector genes form a second-order bipartite network (Figure 3).

Modules 1, 2 and 3: archaeal members of the order Caudovirales

Archaeal members of the order *Caudovirales* form three distinct modules. Module 1 contains haloviruses HVTV-1, HCTV-1 and HCTV-5, all belonging to the family *Siphoviridae* and characterized by large genomes of ~103 kb (5). Module 2 is entirely composed of haloviruses of the family *Myoviridae*. Module 3, the largest of the *Caudovirales* modules, contains 20 viruses, most of these proviruses and members of the family *Siphoviridae*. The two exceptions are the myovirus PhiCh1 and the only known archaeal podovirus HSTV-1.

Modules 1 and 2 each possess many signature genes (58 and 34, respectively) which results from their large genomes and relatively close relatedness. Most of these genes are poorly characterized; exceptions

include a RadA recombinase, a signature of the haloviruses in module 1 and two baseplate proteins (baseplate protein J and spike protein), signatures of the myoviruses in module 2. No signature genes were found for the large and relatively diverse module 3. Instead, module 3 is kept together by a diffuse network of gene sharing and, more importantly, by the set of four essential genes (HK97-like MCP, large subunit of the terminase, portal protein, and capsid maturation protease) which are hallmarks of viruses in the order *Caudovirales* and are shared with viruses from the other two modules as well as tailed bacteriophages (and accordingly do not qualify as signatures of module 3).

The three *Caudovirales* modules encompass most of the genes that are involved in interconnections in the second-order, supermodule network (Figure 3). From that perspective, they are more closely related to each other than any other group of modules in the network. As mentioned above, all three modules are connected by the hallmark genes that make up the morphogenetic toolkit of *Caudovirales*. Additionally, modules 1 and 2 share 7 other genes, including a primase, a nuclease of the Cas4 superfamily and a RNA-primed DNA polymerase of the B family. Modules 2 and 3 are also connected by a tyrosine recombinase which is present, although less commonly, in some plasmids and proviruses outside of the *Caudovirales*.

Module 4: Ligamenvirales

All members of the order *Ligamenvirales* (families *Rudiviridae* and *Lipothrixviridae*) are grouped in module 4. Three signature genes were detected for this module: the unique, four-helix bundle MCP, a SAM-dependent methyltransferase and a glycosyltransferase. The distinct glycosyltransferase connects this module to the ampullaviruses of module 5 and to the only known member of the family *Spiraviridae* (see below). Members of this module also harbor ribbon-helix-helix (RHH) DNA-binding domain proteins (assigned to module 8) which are extensively shared by many archaeal viruses (3).

351 *Module 5: Ampullaviridae*

352 Module 5 encompasses ampullaviruses and family 1 casposons, a recently discovered class of self-
353 synthesizing DNA transposons which employ the Cas1 endonuclease for integration (59-61). The
354 ampullaviruses and the casposons, respectively, comprise two distinct submodules within this module; the
355 two submodules cluster together in 55% of replicas (Figure 2a). The submodules are kept together by the
356 protein-primed DNA PolB which in the context of archaeal viruses is exclusive to this module as well as
357 halophilic viruses His1 and His2 (62). The ampullavirus submodule contains 26 shared genes most of which
358 are refractory to functional annotation. As mentioned above, a glycosyltransferase connects
359 ampullaviruses with members of the *Ligamenvirales*. In addition, a detailed analysis of the sequence of the
360 functionally uncharacterized core proteins led to the identification of two DNA-binding proteins, one
361 containing a winged helix-turn-helix (wHTH) DNA-binding domain and the other an RHH domain (see
362 Supplementary Table S2). The latter protein provides connections to other archaeal virus modules (Figure
363 3).

364

365 *Module 6: Turriviridae and Sphaerolipoviridae*

366 A close analysis of this module reveals a substructure with three submodules that cluster together in 70-
367 90% of replicates. Those submodules consist of (i) *Sphaerolipoviridae*, (ii) *Turriviridae* and related
368 plasmids/proviruses, and (iii) two plasmids related to betasphaerolipovirus SNJ1 (*Halorubrum*
369 *saccharovorum* plasmid pZMX101 and *Methanosarcina acetivorans* plasmid pC2A). With the exception of
370 the latter submodule, all other genomes in module 6 contain the A32-like packaging ATPase, which is a
371 signature gene of this module within the context of archaeal viruses. Instead, pZMX101 and pC2A join the
372 module by their connection to the *Natrinema* sphaerolipovirus SNJ1 through the rolling-circle replication
373 initiation protein RepA (63). Other remarkable genes from this module are the respective MCPs of

374 sphaerolipoviruses and turriviruses, which will be discussed below because of their similarity with the
375 capsid proteins of some bacterial and eukaryotic viruses.

376

377 *Modules 7 and 8: Fuselloviridae and related spindle-shaped viruses*

378 Our analysis provides further clarity on the relationships within the highly divergent group of SSV1-like
379 spindle-shaped viruses. These viruses are split into modules 7 and 8. Module 7 includes *Methanococcus*
380 *voltae* A3 provirus A3-VLP, Pyrococcus abyssi virus 1 (PAV1) and three *Thermococcus* plasmids related to
381 PAV1 (see below). Module 8 includes all classified members of the family *Fuselloviridae* as well as two
382 unclassified spindle-shaped viruses, Aeropyrum pernix spindle-shaped virus 1 and *Thermococcus prieurii*
383 virus 1 (34, 64). Unexpectedly, the only known representative of the family *Guttaviridae*, Aeropyrum pernix
384 ovoid virus 1 (APOV1) (64), is also confidently assigned to this module. The only spindle-shaped virus that is
385 not included in either module 7 or module 8 is salterprovirus His1 (62) which is ambiguously assigned to
386 module 5. Two signature genes are associated with module 7, a putative primase-polymerase and a coiled-
387 coil domain protein. Only the plasmids related to PAV1 contain both signature genes; PAV1 lacks the
388 primase-polymerase whereas A3-VLP lacks the coiled-coil domain protein. The main fusellovirus module 8
389 contains no signature genes but encompasses genes that connect it to the spindle-shaped viruses from
390 module 7 (the fusellovirus MCP and a Zn finger protein that is present in some PAV1-related plasmids) and
391 to the bicaudaviruses from module 9 (DnaA-like AAA+ ATPase and RHH domain protein). The only member
392 of the family *Guttaviridae*, APOV1, is assigned to module 8 through the DnaA-like AAA+ ATPase and an
393 integrase typical of fuselloviruses but it lacks a detectable homologue of the fusellovirus MCP.

394

395 *Module 9: Bicaudaviridae and related single-tailed viruses*

396 Module 9 encompasses crenarchaeal viruses with large spindle-shaped virions that are decorated with one
397 or two long, tail-like appendages protruding from the pointed virion ends. The module includes *Acidianus*
398 two-tailed virus, the only classified member of the family *Bicaudaviridae*, as well as 6 unclassified viruses
399 and three viral genomes assembled from metagenomic data (Sulfolobales virus YNP1, Sulfolobales virus
400 YNP2, and Hyperthermophilic Archaeal Virus 2; Table S1) (27, 40-44). There are four signature genes in this
401 module including a putative integrase and the bicaudavirus MCP (not detected in metagenomic assemblies)
402 which is unrelated to the capsid protein of the smaller spindle-shaped viruses from modules 7 and 8 (32).
403 Apart from the connections with fuselloviruses (see above), some members of this module share a MoxR-
404 like ATPase with the haloviruses from module 1. Metagenomic assemblies, Sulfolobales virus YNP1 and
405 Sulfolobales virus YNP2, harbor two of the signature genes of this module (the putative integrase and a
406 protein with the conserved domain PHA02732, exemplified by the ORF52 of STSV2), as well as the RHH
407 domain protein, the DnaA-like AAA+ ATPase and two uncharacterized proteins present in some other
408 members of the module. Hyperthermophilic Archaeal Virus 2 (40) is assigned to this module based on a
409 single uncharacterized gene (exemplified by the ORF38 of STSV2), which is a signature of bicaudaviruses.

410

411 *Module 10: Pleolipoviridae*

412 Viruses of the family *Pleolipoviridae* have either ssDNA or dsDNA genomes (24) and cluster together in
413 module 10 which encompasses four signature genes. These encode both major structural proteins of
414 pleolipoviruses (the spike protein exemplified by protein VP4 of Halorubrum pleomorphic virus 1 [HRPV-1],
415 and the highly divergent integral membrane protein exemplified by VP3 of HRPV-1), an AAA+ ATPase that
416 has been identified as a virion component in HRPV-1 (but not in other pleolipoviruses), and an
417 uncharacterized protein (GI:226596535). The His2 virus, the sole member of the genus

418 *Gammapleolipovirus*, encodes a pPolB which connects it to module 5 but is nevertheless unambiguously
419 assigned to the pleolipovirus module on the basis of the rest of its core genes.

420

421 *Orphan genomes and ambiguous assignations*

422 In addition to the aforementioned families *Clavaviridae* and *Tristromaviridae*, several viral genomes,
423 despite being connected to the network, cannot be reliably classified into any of the modules. That is, for
424 instance, the case of the only member of the family *Spiraviridae* which is ambiguously assigned to modules
425 3 and 4 although it lacks any of the signature genes of these modules. The spiravirus has only two core
426 genes: an integrase of the tyrosine recombinase superfamily, assigned to the *Caudovirales* module but also
427 widespread in other groups of viruses, and a glycosyltransferase which is shared *Ligamenvirales* and
428 *Ampullaviridae*. Apparently, this genome represents a distinct viral group that does not fit any of the
429 network modules. Indeed, the spiravirus occupies a unique position in the archaeal virosphere in that it is
430 the only known hyperthermophilic virus with an ssDNA genome which is far the largest among all known
431 ssDNA virus genomes (31).

432 A similar situation occurs with the globulovirus Pyrobaculum spherical virus (PSV) which only shares an RHH
433 domain with the rest of the network. As mentioned above, PSV shares multiple genes with TTSV, and
434 addition of such genes to the list of core genes would have resulted in a differentiated Globulovirus
435 module. A third example involves Hyperthermophilic archaeal virus 1 (assembled from metagenomic
436 sequences) which has a single core gene, a glycosyltransferase shared with *Ampullaviridae* and
437 *Ligamenvirales*. In general, highly divergent groups of viruses with only a single or a few available genomes
438 are susceptible to unreliable module assignment, either because they fail to connect with the rest of the
439 network or because they are spuriously assigned a module based on a single, poorly informative gene.

The case of His1 virus is somewhat different because it is ambiguously assigned to the ampullavirus and fusellovirus modules by virtue of two relevant genes, pPolB and the fusellovirus MCP, respectively. Although among the archaeal viruses pPolB is a signature of ampullaviruses, the broader presence of this gene in other bacterial and eukaryotic viruses, which contrasts the exclusivity of the fusellovirus MCP, suggests that His1 virus should be assigned to the fusellovirus module. In addition to those genes, His1 shares the DnaA-like AAA+ ATPase with fuselloviruses and bicaudaviruses, and a glycosyltransferease of the GT-B superfamily with ampullaviruses and members of the *Ligamenvirales*.

The supermodular structure of the network

To explore the existence of hierarchical structure of the archaeal virus network, we applied the module detection algorithm to the second-order bipartite network composed of modules and connector genes. As a result, 5 supermodules were identified (Fig 2 B,C): (i) the *Caudovirales* supermodule that encompasses modules 1, 2 and 3 appears in 68% of the replicas; (ii) modules 4 (*Ligamenvirales*) and 5 (*Ampullaviridae*) form a supermodule in 86% of the replicas; and (iii) modules 7, 8 and 9, which include fuselloviruses and bicaudaviruses, merge in 79% of the replicas. As mentioned above, the *Caudovirales* supermodule is held together by the set of hallmark genes responsible for virion morphogenesis. The *Ligamenvirales* and *Ampullaviridae* modules are linked through a single gene which encodes a glycosyltransferase. The two fusellovirus modules connect through the fusellovirus MCP and a Zn finger protein whereas the largest of these modules connects to the bicaudavirus module 9 through the DnaA-like AAA+ ATPase and RHH domain proteins (also shared with other archaeal viruses, especially those in module 4). Although some turriviruses and pleolipoviruses possess RHH domain proteins, modules 6 and 10 as a whole lack significant connections with the rest of the network and remain unmerged.

462 The network supermodules were used to formalize the intuitive notion of hallmark genes as those genes
463 that are central to a supermodule. Specifically, hallmark genes must be connector genes and appear with
464 high prevalence in at least one supermodule (see Methods). According to these criteria, the archaeal virus
465 network contains 10 hallmark genes: RHH domain protein, large subunit of the terminase, HK97-like MCP,
466 caudoviral prohead protease (U9/U35), portal protein, tyrosine recombinase, DnaA-like AAA+ ATPase,
467 glycosyltransferase, phage Mu protein F and fusellovirus MCP. Note that the MCPs of the *Caudovirales* and
468 of the fuselloviruses are the only capsid proteins that made the list of hallmark genes because those are
469 the only high-level virus taxa that split between more than one primary module.

470 There is one important caveat with regard to the relevance of the supermodules: despite the fact that the
471 supermodules seem to be well supported by their robustness, the supermodular structure of the module-
472 connector gene bipartite network as a whole is not significantly different from the structure of a random
473 network ($p = 0.285$ when comparing the value of Barber's modularity with 200 random networks with the
474 same degree distribution). This lack of statistical significance of the supermodular structure is probably due
475 to the small number of connector, in particular hallmark, genes which makes it difficult to evaluate using
476 topological criteria only whether the modules in the archaeal virus network are actually arranged in
477 supermodules. Instead, the relevance of these supermodules has to be assessed based on biological
478 criteria. More specifically, for each pair of modules, it is important to evaluate the legitimacy of merging
479 based on the particular connector genes they share. For example, there is a substantial body of structural,
480 biochemical and comparative genomic data suggesting that all members of the *Caudovirales* have emerged
481 from a common ancestor (5, 8, 65, 66), supporting the consolidation of modules 1, 2 and 3 into a single
482 supermodule. By contrast, members of the *Ligamenvirales* and *Ampullaviridae* share neither architectural
483 similarity nor clear commonalities in the mode of genome replication. The only gene that brings the two
484 modules together encodes for the glycosyltransferase which likely mediates certain aspects of virus-host
485 interactions and, as is often the case with genes in this functional category, could be independently

acquired from the host by the respective ancestors of the two virus groups or else transferred horizontally between viruses of the two groups. Indeed, besides rudiviruses, lipothrixviruses and ampullaviruses, divergent glycosyltransferases are also encoded by turriviruses, the spiravirus, tristromaviruses and salterprovirus His1. Notably, with the exception of His1, all of these viruses infect hyperthermophilic hosts. Thus, glycosyltransferases might confer advantage to viruses in hot environments. Accordingly, this supermodule appears to reflect common functional features of the constituent viruses. In the third supermodule, the unification of the two groups of fuselloviruses seems to be strongly justified by common structure, genome architecture and gene composition. However, the inclusion of bicaudoviruses could be more on the spurious side, being supported by the promiscuous ATPase and RHH protein genes.

Connections between archaeal viruses and other dsDNA viruses

To gain further insight into the relationship between archaeal and bacterial viruses, we constructed and analyzed a network that contains all available genomes from viruses belonging to the order *Caudovirales*, regardless of the bacterial or archaeal host. In the joint *Caudavirales* network, genomes from the former archaeal modules 1 and 2 again cluster in separate modules, whereas the archaeal viruses from the former module 3 form a larger module together with Phi31-like bacteriophages and numerous unclassified *Siphoviridae*. The latter group of genomes (denoted as 9c in (47)) is itself part of a massive community (module 9 in (47)) that includes lambdoid phages. This community is characterized by intensive gene exchange and a temperate life style. As it occurred with the archaeal module 3, this larger community lacks signature genes and encompasses, instead, the hallmark genes involved in virion morphogenesis as well as the integrase. In accordance with their taxonomy, archaeal viruses from module 2 share several baseplate proteins with bacteriophages of the family *Myoviridae*. Notably, such *Myoviridae*-specific genes appear as

signatures of the archaeal module 2 in the archaeal-only network, but they become connector genes in the complete *Caudovirales* network as the *Myoviridae* split in more than one distinct module.

In a less prominent manner, the archaeal virus network also has connections to eukaryotic viruses and bacteriophages that encode double jelly-roll MCPs. Specifically, the protein-primed PolB found in ampullaviruses also appears in bacterial viruses of the *Tectiviridae*, eukaryotic *Adenoviridae*, virophage Mavirus of the family *Lavidaviridae* and putative viruses-transposons of the Polinton/Maverick (polintovirus) superfamily (20). Moreover, sphaerolipoviruses and turriviruses (module 6) share the A32-like genome packaging ATPase with members of the *Corticoviridae*, *Tectiviridae*, *Adenoviridae*, *Lavidaviridae*, Polintons and the "Megavirales"; the same group of bacterial and eukaryotic viruses shares with turriviruses the double jelly-roll fold MCP and the single jelly-roll minor capsid protein (Figure 3).

Connections between archaeal viruses and non-viral MGE

Our dataset included 14 non-viral MGE: 11 plasmids and 3 casposons. The automatic module detection approach used here placed these elements into modules together with *bona fide* viruses, recapitulating previous observations based on conventional comparative genomics analyses. The non-viral MGE were ascribed to 5 of the 10 defined modules and were connected to the constituent viral genomes primarily via genes encoding the major genome replication proteins. Family 1 casposons integrated in the genomes of Thaumarchaeota encode pPolB (67) and are included into module 5 together with ampullaviruses and salterprovirus His1. Module 6 includes two small plasmids, *Halorubrum saccharovorum* plasmid pZMX101 and *Methanosarcina acetivorans* plasmid pC2A, which share a distinct rolling circle replication initiation endonuclease (RCRE), and by inference, the replication mechanism, with betasphaerolipovirus SNJ1 (63). This module also includes two larger plasmids from *Pyrobaculum oguniense* TE7 and *Thermococcus nautili* (plasmid pTN3); however, given that both of these plasmids encode the DJR MCP and A32-like genome

packaging ATPase, two viral hallmark proteins, as well as some additional viral proteins (22, 23), it appears more likely that these genomes belong to (possibly, defective) proviruses rather than plasmids.

Module 7 includes 3 thermococcal plasmids which collectively share 6 genes with PAV1 including those for several DNA-binding proteins and an AAA+ ATPase (33, 68). It has been hypothesized that more than half of the PAV1 genome has been acquired from plasmids, whereas the remaining portion of the genome has been inherited from spindle-shaped viruses infecting members of the archaeal order Thermococcales (68).

Module 8 includes 2 pRN1-related plasmids, pSSVi and pSSVx. The two plasmids are satellites of fuselloviruses and are involved in a peculiar relationship with the latter (38). Although the plasmids do not encode any structural proteins, upon coinfection with fuselloviruses SSV1 or SSV2, both are encapsidated into spindle-shaped particles which are smaller than the native virions (69, 70). As a result, the plasmids can spread in the host population in a virus-like fashion. Interestingly, pSSVi and pSSVx plasmids each share with fuselloviruses two different genes. pSSVi is included into the module via genes encoding the SSV1-like integrase and an RHH-domain protein, whereas pSSVx encodes the fuselloviral DnaA-like ATPase and an uncharacterized coiled-coil protein conserved in fuselloviruses and exemplified by the A153 protein of SSV1.

Finally, Module 10 includes two small rolling-circle plasmids, *Archaeoglobus profundus* plasmid pGS5 (71) and *Thermococcus prieurii* plasmid pTP2 (72), which connect to members of the genus *Alphapleolipovirus* (Halorubrum pleomorphic viruses 1, 2, and 6) through an RCRE. The latter protein is also shared with the putative provirus MVV which is related to *Turriviridae* from module 6 as well as the monocaudavirus SMV2 from module 9. Notably, the RCRE of SMV2 (YP_009219263) appears to be inactivated: (i) the conserved Motif 2 (HUH, where U = hydrophobic) is changed to YLH; (ii) all homologs of SMV2 RCRE contain two catalytic Tyr residues in Motif 3 (YxxxY, x = any amino acid), a signature of superfamily 1 enzymes (73), whereas SMV2 contains only one of the two tyrosines (YVTKN); (iii) the gene is not conserved in any of the

other bicauda-/monocaudaviruses. Furthermore, the RCRE gene is embedded within genomic neighborhood including several genes encoding proteins annotated as "conjugative plasmid proteins". Thus, it appears that RCRE has been inactivated following its introduction into the SMV2 genome by horizontal gene transfer from a plasmid. The examples presented above clearly demonstrate that the unique archaeal virosphere, at least partially, was shaped by recombination between various selfish replicons, including viruses, plasmids and transposons.

Discussion

Different from their cellular hosts, viruses and related mobile elements lack universal genes (74, 75). As a result, it is often challenging to accurately demonstrate evolutionary connections between distantly related groups of viruses. Indeed, as of now, the highest rank in virus classification is that of order, whereas higher ranks, such as classes or phyla, are not defined due to the absence of obvious marker genes suitable for traditional phylogenetic approaches (76). Although for some large groups of viruses such markers eventually could be defined through further analysis of sequences and structures, network analysis approaches, such as the one described herein, might provide a complementary and perhaps more comprehensive account of the deep evolutionary connections within the viral world and can be useful for guiding the higher-level virus taxonomy.

Bipartite network analysis of the archaeal virosphere revealed 10 distinct modules which generally coincide with the established virus taxonomy and cover 12 different virus families, whereas four additional families remained disconnected from the rest of the virus network. Several unclassified viruses found home within modules containing previously classified viruses, specifically members of the families *Fuselloviridae* and *Bicaudaviridae*, providing a framework for their future classification. The 10 modules display substantial

577 heterogeneity in terms of genomic relatedness and propensity to gene exchange among different groups of
578 viruses. Some modules harbor numerous signature genes, whereas others have few or none. The latter are
579 typically held together by a dense network of shared genes with patchy distribution within the module
580 and/or by highly prevalent core genes shared with other modules. Most of the modules are linked via
581 connector genes encoding a small set of widespread proteins, most notably the RHH-domain containing
582 transcription factors and glycosyltransferases, neither of which is a hallmark viral gene. It appears more
583 likely that the two genes have been independently acquired from the hosts by viruses within each module
584 or spread between viruses horizontally. Such lack of strong connectivity among the modules, with the
585 exception of *Caudovirales* (modules 1-3) and to a lesser extent spindle-shaped viruses (modules 7 and 8),
586 indicates that most of the viral groups within the archaeal virosphere are evolutionarily distinct. In a stark
587 contrast, 5 of the 10 modules include capsidless MGE, suggesting that gene flow between the *bona fide*
588 viruses and such elements played a key role in molding the archaeal virosphere. In this respect, origin and
589 evolution of archaeal viruses mirror that of the eukaryotic virosphere where connections between viruses
590 and various capsid-less elements involve all major groups of viruses and encompass multiple transitions
591 from capsid-less elements to bona fide viruses and vice versa (74, 77, 78). Importantly, apart from the
592 modules including *Caudovirales* and *Turriviridae*/*Sphaerolipoviridae*, archaeal viruses do not display robust
593 evolutionary connections to eukaryotic or bacterial viruses. This is particularly true for viruses infecting
594 hyperthermophilic crenarchaea which continue to occupy a unique position within the global virosphere
595 (47).

596 The observation that non-viral MGE and archaeal viruses are primarily connected through replication
597 proteins could be of particular significance for understanding the origins of the archaeal virosphere. All viral
598 genomes encompass two major components, namely, determinants for virion formation and those for
599 genome replication. In this context, genome replication modules of some archaeal viruses could be derived
600 from different groups of plasmids and, in the case of protein-primed family B DNA polymerases, from self-

601 synthesizing transposons, the Family 1 casposons. By contrast, replication protein genes in other groups of
602 archaeal viruses have been clearly acquired from the host. It has been shown previously that archaeal
603 viruses (and plasmids) from different families and infecting taxonomically distant hosts have acquired the
604 genes for replicative MCM helicases from their respective hosts on multiple independent occasions (79).
605 Interestingly, archaeal members of the *Caudovirales* from module 1 encode nearly complete archaea-
606 specific suites of genome replication proteins. For example, HVTV-1 encodes a DNA polymerase, DNA
607 clamp and its loader, AEP primase and RNase HI (80). Notably, however, some crenarchaeal viruses do not
608 encode any identifiable replication proteins and might therefore employ unique genome replication
609 strategies or encode specific proteins for hijacking the host replication machinery. The origin of the other
610 major component of the viral genomes, namely determinants for virion structure, is more difficult to trace.
611 By definition, structural proteins encoded by archaea-specific viruses have no homologs among bacterial
612 and eukaryotic viruses. Nevertheless, we have recently described a case where one of the major
613 nucleocapsid proteins of tristromavirus TTV1 has been exapted from the inactivated Cas4-like nuclease
614 (81). Thus, the replication module in many archaeal viruses can be traced to non-viral MGE or the archaeal
615 hosts, whereas structural proteins of archaeal viruses (and viruses in general) can occasionally evolve from
616 cellular proteins that have no *a priori* role in virion formation.

617 The results of the network analysis of archaeal viruses differ from those previously reported for viruses of
618 bacteria and eukaryotes (47) in that the modules of the archaeal network, with the exception of the
619 *Caudovirales*, are quite sparsely connected; so much so that although supermodules were identified, their
620 reality could not be supported statistically. Overall, only 9% of the connections in the archaeal network
621 involve members of different modules (excluding the *Caudovirales*), whereas such intermodule
622 connections constitute 25% of the bacterial *Caudovirales* network. The explanation of this sharp distinction
623 is likely to be twofold. First, the current sampling of archaeal viruses is likely to be much less representative
624 of their true diversity than the sampling of viruses of bacteria and eukaryotes. Nearly all hyperthermophilic

archaea possess CRISPR-Cas adaptive immunity loci, often multiple ones (82), which is indicative of the perennial coevolution of these archaea with diverse viromes. Yet, viruses of archaeal hyperthermophiles outside Crenarchaeota remain virtually unknown. Second, the paucity of connections in the archaeal network could reflect actual different origins of the distinct groups of archaeal viruses, in particular from different non-viral MGE. Further, increasingly extensive exploration of the archaeal virosphere should elucidate the relative contributions of these two factors to the architecture of the viral network.

Funding information

JI and EVK are supported by intramural funds of the US Department of Health and Human Services (to the National Library of Medicine). DP was supported by the Agence Nationale de la Recherche (ANR) (program BLANC project EXAVIR).

637 **References**

638 1. **Pietilä MK, Demina TA, Atanasova NS, Oksanen HM, Bamford DH.** 2014. Archaeal viruses and
639 bacteriophages: comparisons and contrasts. Trends Microbiol **22:**334-344.
640 2. **Prangishvili D.** 2013. The wonderful world of archaeal viruses. Annu Rev Microbiol **67:**565-585.
641 3. **Prangishvili D, Garrett RA, Koonin EV.** 2006. Evolutionary genomics of archaeal viruses: unique
642 viral genomes in the third domain of life. Virus Res **117:**52-67.
643 4. **Snyder JC, Bolduc B, Young MJ.** 2015. 40 Years of archaeal virology: Expanding viral diversity.
644 Virology **479-480:**369-378.
645 5. **Sencilo A, Jacobs-Sera D, Russell DA, Ko CC, Bowman CA, Atanasova NS, Osterlund E, Oksanen**
646 **HM, Bamford DH, Hatfull GF, Roine E, Hendrix RW.** 2013. Snapshot of haloarchaeal tailed virus
647 genomes. RNA Biol **10:**803-816.
648 6. **Atanasova NS, Bamford DH, Oksanen HM.** 2016. Virus-host interplay in high salt environments.
649 Environ Microbiol Rep **8:**431-444.
650 7. **Pfister P, Wasserfallen A, Stettler R, Leisinger T.** 1998. Molecular analysis of Methanobacterium
651 phage psiM2. Mol Microbiol **30:**233-244.
652 8. **Krupovic M, Forterre P, Bamford DH.** 2010. Comparative analysis of the mosaic genomes of tailed
653 archaeal viruses and proviruses suggests common themes for virion architecture and assembly
654 with tailed viruses of bacteria. J Mol Biol **397:**144-160.
655 9. **Krupovic M, Spang A, Gribaldo S, Forterre P, Schleper C.** 2011. A thaumarchaeal provirus testifies
656 for an ancient association of tailed viruses with archaea. Biochem Soc Trans **39:**82-88.
657 10. **Luo Y, Pfister P, Leisinger T, Wasserfallen A.** 2001. The genome of archaeal prophage PsiM100
658 encodes the lytic enzyme responsible for autolysis of Methanothermobacter wolfeii. J Bacteriol
659 **183:**5788-5792.

11. **Pawlowski A, Rissanen I, Bamford JK, Krupovic M, Jalasvuori M.** 2014. Gammasphaerolipovirus, a newly proposed bacteriophage genus, unifies viruses of halophilic archaea and thermophilic bacteria within the novel family Sphaerolipoviridae. Arch Virol **159:**1541-1554.

12. **Bamford DH, Ravantti JJ, Ronnholm G, Laurinavicius S, Kukkaro P, Dyall-Smith M, Somerharju P, Kalkkinen N, Bamford JK.** 2005. Constituents of SH1, a novel lipid-containing virus infecting the halophilic euryarchaeon Haloarcula hispanica. J Virol **79:**9097-9107.

13. **Porter K, Tang SL, Chen CP, Chiang PW, Hong MJ, Dyall-Smith M.** 2013. PH1: an archaeovirus of Haloarcula hispanica related to SH1 and HHIV-2. Archaea **2013:**456318.

14. **Jalasvuori M, Jaatinen ST, Laurinavicius S, Ahola-Iivarinen E, Kalkkinen N, Bamford DH, Bamford JK.** 2009. The closest relatives of icosahedral viruses of thermophilic bacteria are among viruses and plasmids of the halophilic archaea. J Virol **83:**9388-9397.

15. **Zhang Z, Liu Y, Wang S, Yang D, Cheng Y, Hu J, Chen J, Mei Y, Shen P, Bamford DH, Chen X.** 2012. Temperate membrane-containing halophilic archaeal virus SNJ1 has a circular dsDNA genome identical to that of plasmid pHH205. Virology **434:**233-241.

16. **Gil-Carton D, Jaakkola ST, Charro D, Peralta B, Castano-Diez D, Oksanen HM, Bamford DH, Abrescia NG.** 2015. Insight into the assembly of viruses with vertical single beta-barrel major capsid proteins. Structure **23:**1866-1877.

17. **Rissanen I, Grimes JM, Pawlowski A, Mantynen S, Harlos K, Bamford JK, Stuart DI.** 2013. Bacteriophage P23-77 capsid protein structures reveal the archetype of an ancient branch from a major virus lineage. Structure **21:**718-726.

18. **Veesler D, Ng TS, Sendamarai AK, Eilers BJ, Lawrence CM, Lok SM, Young MJ, Johnson JE, Fu CY.** 2013. Atomic structure of the 75 MDa extremophile Sulfolobus turreted icosahedral virus determined by CryoEM and X-ray crystallography. Proc Natl Acad Sci U S A **110:**5504-5509.

19. **Krupovic M, Bamford DH.** 2008. Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol **6:**941-948.

20. **Krupovic M, Koonin EV.** 2015. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nat Rev Microbiol **13:**105-115.

21. **Krupovic M, Bamford DH.** 2008. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology **375:**292-300.

22. **Gaudin M, Krupovic M, Marguet E, Gauliard E, Cvirkaite-Krupovic V, Le Cam E, Oberto J, Forterre P.** 2014. Extracellular membrane vesicles harbouring viral genomes. Environ Microbiol **16:**1167-1175.

23. **Bernick DL, Karplus K, Lui LM, Coker JK, Murphy JN, Chan PP, Cozen AE, Lowe TM.** 2012. Complete genome sequence of Pyrobaculum oguniense. Stand Genomic Sci **6:**336-345.

24. **Pietilä MK, Roine E, Sencilo A, Bamford DH, Oksanen HM.** 2016. *Pleolipoviridae*, a newly proposed family comprising archaeal pleomorphic viruses with single-stranded or double-stranded DNA genomes. Arch Virol **161:**249-256.

25. **Roine E, Kukkaro P, Paulin L, Laurinavicius S, Domanska A, Somerharju P, Bamford DH.** 2010. New, closely related haloarchaeal viral elements with different nucleic acid types. J Virol **84:**3682-3689.

26. **Peng X, Basta T, Haring M, Garrett RA, Prangishvili D.** 2007. Genome of the Acidianus bottle-shaped virus and insights into the replication and packaging mechanisms. Virology **364:**237-243.

27. **Gudbergsdóttir SR, Menzel P, Krogh A, Young M, Peng X.** 2016. Novel viral genomes identified from six metagenomes reveal wide distribution of archaeal viruses and high viral diversity in terrestrial hot springs. Environ Microbiol **18:**863-874.

28. **Prangishvili D, Krupovic M.** 2012. A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales". Arch Virol **157:**791-795.

29. **Rensen EI, Mochizuki T, Quemin E, Schouten S, Krupovic M, Prangishvili D.** 2016. A virus of hyperthermophilic archaea with a unique architecture among DNA viruses. Proc Natl Acad Sci U S A **113:**2478-2483.

30. **Mochizuki T, Yoshida T, Tanaka R, Forterre P, Sako Y, Prangishvili D.** 2010. Diversity of viruses of the hyperthermophilic archaeal genus Aeropyrum, and isolation of the Aeropyrum pernix bacilliform virus 1, APBV1, the first representative of the family Clavaviridae. Virology **402:**347-354.

31. **Mochizuki T, Krupovic M, Pehau-Arnaudet G, Sako Y, Forterre P, Prangishvili D.** 2012. Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. Proc Natl Acad Sci U S A **109:**13386-13391.

32. **Krupovic M, Quemin ER, Bamford DH, Forterre P, Prangishvili D.** 2014. Unification of the globally distributed spindle-shaped viruses of the Archaea. J Virol **88:**2354-2358.

33. **Geslin C, Gaillard M, Flament D, Rouault K, Le Romancer M, Prieur D, Erauso G.** 2007. Analysis of the first genome of a hyperthermophilic marine virus-like particle, PAV1, isolated from Pyrococcus abyssi. J Bacteriol **189:**4510-4519.

34. **Gorlas A, Koonin EV, Bienvenu N, Prieur D, Geslin C.** 2012. TPV1, the first virus isolated from the hyperthermophilic genus Thermococcus. Environ Microbiol **14:**503-516.

35. **Iverson E, Stedman K.** 2012. A genetic study of SSV1, the prototypical fusellovirus. Front Microbiol **3:**200.

36. **Prangishvili D, Stedman K, Zillig W.** 2001. Viruses of the extremely thermophilic archaeon Sulfolobus. Trends Microbiol **9:**39-43.

37. **Redder P, Peng X, Brugger K, Shah SA, Roesch F, Greve B, She Q, Schleper C, Forterre P, Garrett RA, Prangishvili D.** 2009. Four newly isolated fuselloviruses from extreme geothermal environments reveal unusual morphologies and a possible interviral recombination mechanism. Environ Microbiol **11:**2849-2862.

38. **Contursi P, Fusco S, Cannio R, She Q.** 2014. Molecular biology of fuselloviruses and their satellites. Extremophiles **18:**473-489.

39. **Häring M, Vestergaard G, Rachel R, Chen L, Garrett RA, Prangishvili D.** 2005. Virology: independent virus development outside a host. Nature **436:**1101-1102.

40. **Garrett RA, Prangishvili D, Shah SA, Reuter M, Stetter KO, Peng X.** 2010. Metagenomic analyses of novel viruses and plasmids from a cultured environmental sample of hyperthermophilic neutrophiles. Environ Microbiol **12:**2918-2930.

41. **Erdmann S, Chen B, Huang X, Deng L, Liu C, Shah SA, Le Moine Bauer S, Sobrino CL, Wang H, Wei Y, She Q, Garrett RA, Huang L, Lin L.** 2014. A novel single-tailed fusiform Sulfolobus virus STSV2 infecting model Sulfolobus species. Extremophiles **18:**51-60.

42. **Xiang X, Chen L, Huang X, Luo Y, She Q, Huang L.** 2005. Sulfolobus tengchongensis spindle-shaped virus STSV1: virus-host interactions and genomic features. J Virol **79:**8677-8686.

43. **Erdmann S, Le Moine Bauer S, Garrett RA.** 2014. Inter-viral conflicts that exploit host CRISPR immune systems of Sulfolobus. Mol Microbiol **91:**900-917.

44. **Hochstein RA, Amenabar MJ, Munson-McGee JH, Boyd ES, Young MJ.** 2016. Acidianus Tailed Spindle Virus: a New Archaeal Large Tailed Spindle Virus Discovered by Culture-Independent Methods. J Virol **90:**3458-3468.

45. **Bolduc B, Shaughnessy DP, Wolf YI, Koonin EV, Roberto FF, Young M.** 2012. Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. J Virol **86:**5562-5573.

46. **Held NL, Whitaker RJ.** 2009. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ Microbiol **11:**457-466.

47. **Iranzo J, Krupovic M, Koonin EV.** 2016. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. mBio **7:**e00978-00916.

755 48. **Li W, Godzik A.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or
756 nucleotide sequences. Bioinformatics **22:**1658-1659.

757 49. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST
758 and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389-
759 3402.

760 50. **Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF.** 2001.
761 Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics
762 and other refinements. Nucleic Acids Res **29:**2994-3005.

763 51. **Rosvall M, Bergstrom CT.** 2008. Maps of random walks on complex networks reveal community
764 structure. Proc Natl Acad Sci U S A **105:**1118-1123.

765 52. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput.
766 Nucleic Acids Res **32:**1792-1797.

767 53. **Meier A, Söding J.** 2015. Automatic prediction of protein 3D structures by probabilistic multi-
768 template homology modeling. PLoS Comput Biol **11:**e1004343.

769 54. **Söding J.** 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics **21:**951-
770 960.

771 55. **Makarova KS, Wolf YI, Koonin EV.** 2015. Archaeal Clusters of Orthologous Genes (arCOGs): An
772 Update and Application for Analysis of Shared Features between Thermococcales,
773 Methanococcales, and Methanobacteriales. Life (Basel) **5:**818-840.

774 56. **Barber MJ.** 2007. Modularity and community detection in bipartite networks. Phys Rev E Stat
775 Nonlin Soft Matter Phys **76:**066102.

776 57. **Marquitti FM, Guimaraes PR, Pires MM, Bittencourt LF.** 2014. Modular: software for the
777 autonomous computation of modularity in large network sets. Ecography **37:**221-224.

58. **Bascompte J, Jordano P, Melian CJ, Olesen JM.** 2003. The nested assembly of plant-animal mutualistic networks. Proc Natl Acad Sci U S A **100:**9383-9387.

59. **Krupovic M, Koonin EV.** 2016. Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. Curr Opin Microbiol **31:**25-33.

60. **Béguin P, Charpin N, Koonin EV, Forterre P, Krupovic M.** 2016. Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res **in press**.

61. **Hickman AB, Dyda F.** 2015. The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications. Nucleic Acids Res **43:**10576-10587.

62. **Bath C, Cukalac T, Porter K, Dyall-Smith ML.** 2006. His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. Virology **350:**228-239.

63. **Wang Y, Sima L, Lv J, Huang S, Liu Y, Wang J, Krupovic M, Chen X.** 2016. Identification, characterization, and application of the replicon region of the halophilic temperate sphaerolipovirus SNJ1. J Bacteriol **198:**1952-1964.

64. **Mochizuki T, Sako Y, Prangishvili D.** 2011. Provirus induction in hyperthermophilic archaea: characterization of Aeropyrum pernix spindle-shaped virus 1 and Aeropyrum pernix ovoid virus 1. J Bacteriol **193:**5412-5419.

65. **Krupovic M, Bamford DH.** 2011. Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. Curr Opin Virol **1:**118-124.

66. **Pietilä MK, Laurinmaki P, Russell DA, Ko CC, Jacobs-Sera D, Hendrix RW, Bamford DH, Butcher SJ.** 2013. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. Proc Natl Acad Sci U S A **110:**10604-10609.

67. **Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV.** 2014. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol **12:**36.

68. **Krupovic M, Gonnet M, Hania WB, Forterre P, Erauso G.** 2013. Insights into dynamics of mobile genetic elements in hyperthermophilic environments from five new Thermococcus plasmids. PLoS One **8:**e49044.

69. **Arnold HP, She Q, Phan H, Stedman K, Prangishvili D, Holz I, Kristjansson JK, Garrett R, Zillig W.** 1999. The genetic element pSSVx of the extremely thermophilic crenarchaeon Sulfolobus is a hybrid between a plasmid and a virus. Mol Microbiol **34:**217-226.

70. **Wang Y, Duan Z, Zhu H, Guo X, Wang Z, Zhou J, She Q, Huang L.** 2007. A novel Sulfolobus non-conjugative extrachromosomal genetic element capable of integration into the host genome and spreading in the presence of a fusellovirus. Virology **363:**124-133.

71. **Lopez-Garcia P, Forterre P, van der Oost J, Erauso G.** 2000. Plasmid pGS5 from the hyperthermophilic archaeon Archaeoglobus profundus is negatively supercoiled. J Bacteriol **182:**4998-5000.

72. **Gorlas A, Krupovic M, Forterre P, Geslin C.** 2013. Living side by side with a virus: characterization of two novel plasmids from Thermococcus prieurii, a host for the spindle-shaped virus TPV1. Appl Environ Microbiol **79:**3822-3828.

73. **Ilyina TV, Koonin EV.** 1992. Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. Nucleic Acids Res **20:**3279-3285.

74. **Koonin EV, Dolja VV.** 2014. Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev **78:**278-303.

75. **Koonin EV, Senkevich TG, Dolja VV.** 2006. The ancient Virus World and evolution of cells. Biol Direct **1:**29.

76. **King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ.** 2011. Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego.

77. **Koonin EV, Dolja VV, Krupovic M.** 2015. Origins and evolution of viruses of eukaryotes: The ultimate modularity. Virology **479-480:**2-25.

78. **Krupovic M.** 2013. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol **3:**578-586.

79. **Krupovic M, Gribaldo S, Bamford DH, Forterre P.** 2010. The evolutionary history of archaeal MCM helicases: a case study of vertical evolution combined with hitchhiking of mobile genetic elements. Mol Biol Evol **27:**2716-2732.

80. **Kazlauskas D, Krupovic M, Venclovas C.** 2016. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. Nucleic Acids Res **44:**4551-4564.

81. **Krupovic M, Cvirkaite-Krupovic V, Prangishvili D, Koonin EV.** 2015. Evolution of an archaeal virus nucleocapsid protein from the CRISPR-associated Cas4 nuclease. Biol Direct **10:**65.

82. **Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJ, Charpentier E, Haft DH, Horvath P, Moineau S, Mojica FJ, Terns RM, Terns MP, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R, Koonin EV.** 2015. An updated evolutionary classification of CRISPR-Cas systems. Nat Rev Microbiol **13:**722-736.

844 **Figure legends**

845 **Figure 1: The bipartite network of archaeal viruses.** Archaeal genomes are represented as colored circles,
846 genes are denoted by the intersections of edges. The color of a genome node accords with its module
847 assignation. To provide a wider context to the archaeal virus network, tailed bacteriophages of the order
848 *Caudovirales* are shown in gray, whereas the two bacterial sphaerolipoviruses are shown in white. Nonviral
849 mobile genetic elements, including plasmid and casposons, which are connected to different viral modules
850 are represented as triangles. Edges involving connector genes are colored in black, whereas those involving
851 signature genes are in the same colors are their respective modules. The genomes of *Tristromaviridae* and
852 *Clavaviridae* do not harbor any core genes, and therefore they appear disconnected from the rest of the
853 network. GT: glyclosyltransferase of the GT-B superfamily, RHH: ribbon-helix-helix domain-containing
854 protein, pPolB: protein-primed DNA polymerase B, MCP: major capsid protein, HAV1: Hyperthermophilic
855 Archaeal Virus 1.

856

857 **Figure 2: Robustness and cross-similarities of modules (A, B) and supermodules (C, D) in the archaeal**
858 **virus bipartite network.** The module detection algorithm was run in 100 replicas of the original network,
859 yielding 100 alternative partitions of the network. Of those partitions, the one with the highest value of the
860 Barber's modularity index was selected as the optimal partition; the other 99 were used to assess the
861 robustness of the modules in the optimal partition. A,C: the heat map represents the average fraction of
862 replicas in which a pair of genomes was grouped in the same module. Genomes are sorted in both axes
863 based on the module they belong to in the optimal partition. B, D: Same for gene families. Dark blocks
864 correspond to robust modules, with size proportional to the number of genomes or gene families in the
865 module. Lighter shading within a block suggests the existence of internal structure, while shaded regions
866 between blocks are indicative of supermodular structure. The asterisk in A denotes the ambiguous

867 assignation of His1 virus to modules 5 and 8. See main text and Table 1 for a description of the contents of
868 each module.

869

870 **Figure 3: Second-order structure of the archaeal virus network.** Large circles represent modules, with their
871 size proportional to the number of genomes they encompass. Black dots represent connector genes, i.e.
872 genes whose prevalence in two modules is greater than exp(-1). Light gray edges have been used to
873 indicate the occasional presence of a connector gene in an otherwise disconnected module. MCP: major
874 capsid protein, DJR MCP: double jelly-roll fold major capsid protein, RHH: ribbon-helix-helix domain-
875 containing protein, pPolB: protein-primed DNA polymerase B.

876

877

878 **Table 1: Modules in the archaeal virus network.** The robustness of a module is the average fraction of replicas in which pairs of members of that
879 module are grouped together. The distinctiveness of a module is the average fraction of replicas in which members of that module are only
880 grouped with members of the same module. A low value of distinctiveness for a module is indicative of its belonging to a larger supermodule.
881 MCP: major capsid protein. The density is the fraction of connections relative to all possible gene-genome pairs in a module. Low density
882 indicates module heterogeneity, which may be intrinsic to the module (e.g. in module 3) due to the existence of submodules (e.g. in modules 4,
883 5 and 6).

	Norg	Robust. Org	Distinct. Org	Nfam	Robust. Fam	Distinct. Fam	Density	Composition (genomes)	Composition (gene families)
1	3	1.00	1.00	63	1.00	0.99	1.00	Haloviruses HVTV-1, HCTV-1, HCTV-5	RadA recombinase
2	7	0.96	0.44	43	0.98	0.84	0.86	*Myoviridae*	Baseplate protein J, baseplate spike
3	20	0.92	0.84	15	0.86	0.57	0.42	Other members of the *Caudovirales*	Large subunit of the terminase, HK97-like MCP, protease (U9/U35), integrase, portal protein
4	14	0.99	0.97	39	0.98	0.99	0.45	*Lipothrixviridae*, *Rudiviridae*	MCP from viruses of the order *Ligamenvirales*, glycosyltransferase, SAM-dependent methyltransferase
5	7	0.72	0.70	31	0.88	0.96	0.45	*Ampullaviridae*, family 1 casposons	Protein-primed DNA PolB
6	14	0.83	0.81	30	0.95	0.95	0.26	*Sphaerolipoviridae*, *Turriviridae* (and related)	A32-like packaging ATPase (FtsK/HerA), DJR MCP (only *Turriviridae* and related proviruses)
7	5	1.00	0.62	6	0.89	0.54	0.67	*Pyrococcus abyssi* virus 1, *Methanococcus voltae* A3 provirus A3-VLP and related *Thermococcus* plasmids	Putative primase-polymerase, coiled-coil domain protein
8	18	1.00	0.95	14	1.00	0.96	0.69	Most *Fuselloviridae*, *Guttaviridae*	RHH domain protein, DnaA-like AAA+ ATPase, MCP from fuselloviruses (only *Fuselloviridae*)
9	10	1.00	1.00	27	1.00	0.98	0.47	*Bicaudaviridae*	MoxR-like ATPase, putative integrase, MCP from *Bicaudaviridae*
10	11	0.98	0.88	5	0.98	0.72	0.68	*Pleolipoviridae*	AAA+ ATPase, major spike protein, integral membrane protein (except for His2), uncharacterized protein

884

885 **Table 2: Connector genes**

Fam. Nr	Representative seq. (GIs)	Annotation	Modules with high prevalence
30578	448260172	RHH domain	4, 8, 9
24	506497871	Integrase, tyrosine recombinase superfamily	2, 3
5	33323612	Terminase, large subunit	1, 2, 3
16	340545227	Portal protein	1, 2, 3
13	90110596	Major capsid protein, HK97-like	1, 2, 3
11	738838588	DNA PolB *	1, 2, 5
30596	448260216	DnaA-like AAA+ ATPase (PHA00729)	8, 9
111	9634157	Protease (herpesvirus S21, phage U9/U35)	1, 2, 3
30580	146411830	Glycosyltransferase, GT-B superfamily	4, 5
26	294663759	Phage Mu protein F	1, 3
30576	472438248	Major capsid protein from fuselloviruses	7, 8
27	310831525	HNHc endonuclease	1, 2
60	45686344	Metallophosphatase, MPP superfamily	1, 2
551	22091125	HTH domain	1, 2
30601	270281838	Zinc finger protein	7, 8
305	353228106	MoxR-like ATPase	1, 9
69	294338118	PDDEXK nuclease, Cas4 superfamily	1, 2
6697	156564162	Archaeo-eukaryotic primase	1, 2
19	9628153	Ribonucleotide reductase, large subunit	1, 2

886 * Protein-primed in module 5, RNA-primed in modules 1 and 2.

887