AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Virology

Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements

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

National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland, USA[a]; Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Paris, France[b]

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 similar previously analyzed networks of eukaryotic and bacterial viruses, the archaeal virus network is sparsely connected. With the exception of the tailed viruses related to bacteriophages of the order *Caudovirales* and the families *Turriviridae* and *Sphaerolipoviridae* that are linked to a distinct supermodule of eukaryotic and bacterial viruses, there are few connector genes shared by different archaeal virus modules. In contrast, most of these modules include, in addition to viruses, capsidless mobile elements, emphasizing tight evolutionary connections between the two types of entities in archaea. The relative contributions of distinct evolutionary origins, in particular from nonviral 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 5,740 protein sequences from 116 genomes allowed dissection of the archaeal virus network and showed that most groups of archaeal viruses are evolutionarily connected to capsidless mobile genetic elements, including various plasmids and transposons. This finding could reflect actual independent origins of the distinct groups of archaeal viruses from different nonviral elements, providing important insights into the emergence and evolution of the archaeal virome.

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

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

The recently created family *Sphaerolipoviridae* includes viruses with tailless icosahedral capsids and internal membranes. Members of the genera *Alphasphaerolipovirus* and *Betasphaerolipovirus* infect halophilic archaea, whereas those of the genus *Gammasphaerolipovirus* propagate in bacteria of the genus *Thermus* (11). Alphasphaerolipoviruses possess linear double-stranded DNA (dsDNA) genomes (12, 13), whereas betasphaerolipoviruses and gammas-

Received 17 August 2016 **Accepted** 19 September 2016

Accepted manuscript posted online 28 September 2016

Citation Iranzo J, Koonin EV, Prangishvili D, Krupovic M. 2016. Bipartite network analysis of the archaeal virosphere: evolutionary connections between viruses and capsidless mobile elements. J Virol 90:11043–11055. doi:10.1128/JVI.01622-16.

Editor: R. M. Sandri-Goldin, University of California, Irvine

Address correspondence to Mart Krupovic, krupovic@pasteur.fr.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JVI.01622-16.

phaerolipoviruses encapsidate circular genomes (14, 15). Nevertheless, virion organization and morphogenesis are similar among viruses in all three genera; all these viruses encode two major capsid proteins (MCPs) with a single jelly-roll fold and apparently encapsidate their genomes by using homologous A32-like packaging ATPases (16, 17).

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

Archaea-specific viruses are classified into 10 families (2). Arguably, the most unexpected among these viruses are members of the family *Ampullaviridae* with bottle-shaped virions. This family is currently represented by a single isolate, Acidianus bottle-shaped virus (ABV) (26), but two additional complete ABV-like genomes were recently assembled from metagenomic data (27). Ampullaviruses contain linear dsDNA genomes with terminal inverted repeats, which appear to be replicated by the virus-encoded protein-primed DNA polymerases of the B family (26). Crenarchaea are infected by a range of filamentous viruses, which can be flexible or rigid and long or short and contain dsDNA or ssDNA genomes. These viruses are classified into 5 families: *Rudiviridae*, *Lipothrixviridae*, *Tristromaviridae*, *Clavaviridae*, and *Spiraviridae*. Rod-shaped, nonenveloped rudiviruses and flexible, and 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 the *Tristromaviridae*; these viruses do not share genes with other archaeal viruses and have a virion organization that 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). In contrast, the Acidianus coil-shaped virus (ACV) of the family *Spiraviridae* contains by far the largest genome (~25 kb) among known ssDNA viruses. The ACV virion is organized as a coil that 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 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 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, the *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 crenarchaeal viruses. A global comparative genomic analysis of archaeal viruses, performed a decade ago, 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 nonviral 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 nonviral elements in this process.

MATERIALS AND 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*, *Rudiviridae*, *Sphaerolipoviridae* (note that viruses of the genera *Alphasphaerolipovirus* and *Betasphaerolipovirus* infect archaea, whereas those of the genus *Gammasphaerolipovirus* infect bacteria), *Spiraviridae*, *Tristromaviridae*, and *Turriviridae* as well as members of the order *Caudovirales* (families *Siphoviridae*, *Podoviridae*, and *Myoviridae*) that infect *Archaea*. This data set was complemented with sequences of unclassified archaeal viruses, including those assembled from metagenomic data, as well as previously described proviruses (8, 9, 21, 46), plasmids, and casposons known to share genes with archaeal viruses. In total, the initial data set contained 5,740 protein sequences from 116 genomes.

Classification of genes into homologous families. Following the same methodology as the one described previously (47), all protein sequences were initially clustered at 90% identity and 70% coverage by using CD-HIT (48) to generate a nonredundant data set. For each sequence in this set, a BLASTp search (49) with composition-based statistics (50) and filtering of low-complexity regions was carried out against all other sequences. An E value cutoff equal to 0.01 (database size fixed to 2e7) was used to determine valid hits. The scores for these hits were subsequently collected from a BLASTp search with neither composition-based statistics nor a low-complexity filter. The set of scored BLAST hits defined a weighted sequence similarity network that we partitioned with Infomap (51) (100 trials; 2-level hierarchy) in order to generate preliminary groups of homologous genes. In the next step, we applied profile analysis to find and merge groups of related sequences. For this purpose, sequences in each group were aligned with Muscle (52) (default parameters), and the alignments were used to predict secondary structure and build profiles with the tools "addss" and "hhmake" available in the HH-suite package (53). The collection of profiles was enriched with those generated previously (47) for a large number of (nonarchaeal) dsDNA viruses. Profile-profile comparisons were carried out by using HHsearch (54). To accept or reject hits, we applied the same heuristics as those described previously (47): hits with a probability of >0.90 were accepted if they covered at least 50% of the length of the profile; additionally, hits with a coverage of 20% or greater were also accepted if their probability was >0.99 and their length was >100 amino acids (aa). This pipeline rendered a total of 2,931 clusters of homologous sequences, 938 of which comprised multiple sequences and 1,993 of which 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 the *Ligamenvirales*, and integral membrane proteins 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 a 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 gene loss rates can be easily computed provided that (i) there is a collection of pairs of genomes, with the gene of interest being 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 and N_j are the numbers of families in each genome (47). Relative to this distance, the time from the last common ancestor can be simply expressed as $t = D_{ij}/2$. We then used the pure-loss evolutionary model to estimate the loss rate for 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 time t is $P_1 = e^{-rt}$, 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 of 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 a 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} and P_{11} and the distance 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 horizontal gene transfer (HGT) could bias loss rate estimates; thus, we considered only those pairs of genomes with distances D_{ij} smaller than 1. Genes with a loss rate r smaller than 1 were assigned to the "core." In this way, we obtained a list of 2,560 core gene families, 180 of which were not classified as core gene families in a previous analysis of the dsDNA virus world (47). Such an 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 a significant presence in archaeal viruses. Table S2 in the supplemental material 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 1 implies that the family is present in all genomes of the data set. 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 >90% of their gene content (including 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 the only available genome from a member of the *Clavaviridae* (Aeropyrum pernix bacilliform virus 1), both representatives of the *Tristromaviridae* (Pyrobaculum filamentous virus 1 and Thermoproteus tenax virus 1), and the globulovirus Thermoproteus tenax spherical virus 1 (TTSV), 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 an 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 nonconnector genes whose abundance exceeded the threshold in a single module. We detected supermodules 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 previously (47). To this 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 being a hallmark gene if it fulfilled two conditions: (i) it is a connector gene and (ii) it has a prevalence greater than a given threshold in at least one of the supermodules. Any prevalence threshold of 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 an MI value of >0.6 for the best match and an MI value of <0.02 for the second-best match.

RESULTS

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

To enrich the representation of certain archaeal virus families and to further investigate the evolutionary connections between viruses and nonviral mobile genetic elements (MGEs), we included 16 previously described archaeal proviruses (related to viruses of the order *Caudovirales* and the families *Fuselloviridae*, *Turriviridae*, and *Pleolipoviridae*), 11 archaeal plasmids, and 3 casposons (self-synthesizing transposons). The only member of the family *Clavaviridae* (Aeropyrum pernix bacilliform virus 1) does not share genes with the rest of the archaeal viruses and therefore remains separated from the network. After combining highly similar genomes, the bipartite network of archaeal viruses consisted of 111 genomes and 2,883 gene families.

For efficient analysis of a complex network, it is desirable to minimize the effect of noisy connections that 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 includes only "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 (Fig. 1) includes a giant connected component that contains 107 (pro)viral genomes and 274 core gene families. Apart from the giant component and in addition to the above-mentioned "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 TTSV. Accordingly, these genomes remain isolated from the network. The two tristromaviruses as well as TTSV and the other globulovirus, Pyrobaculum spherical virus (PSV), share genes with each other. However, because the core detection algorithm requires that a gene be present in at least three genomes, none of these shared genes could be classified as core genes. 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 ($P < 0.01$ compared to a random network).

The network consists of 10 robust modules (Table 1; see also Table S1 in the supplemental material), most of which encompass viruses from one or, in some cases, two families (Fig. 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 Materials and Methods; see also Table S2 in the supplemental material). Informally, signature genes are nearly exclusive to a particular module within a 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 (Fig. 1). We defined connector genes as those genes that 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 (Fig. 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 the haloviruses Haloarcula vallismortis tailed virus 1 (HVTV-1), Haloarcula californiae tailed virus 1 (HCTV-1), and HCTV-5, all belonging to the family *Siphoviridae* and characterized by large genomes of ~103 kb (5).

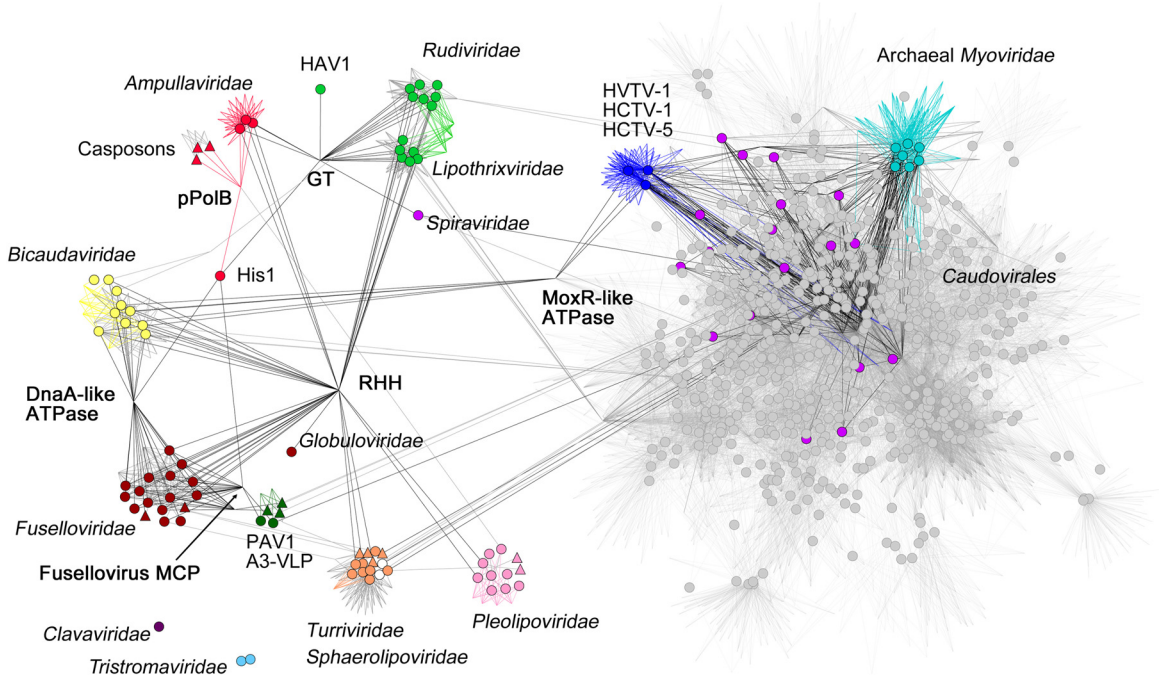

FIG 1 The bipartite network of archaeal viruses. Archaeal genomes are represented as colored circles, and genes are denoted by the intersections of edges. The color of a genome node accords with its module assignation. To provide a wider context to the archaeal virus network, tailed bacteriophages of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

Module 2 is composed entirely of haloviruses of the family *Myoviridae*. Module 3, the largest of the *Caudovirales* modules, contains 20 viruses, most of which are proviruses and members of the family *Siphoviridae*. The two exceptions are the myovirus PhiCh1 and the only known archaeal podovirus, Halorubrum sodomense tailed virus 1 (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 a set of four essential genes (HK97-like MCP, large subunit of the terminase, portal protein, and capsid maturation protease), which are hallmarks of viruses of 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 (Fig. 3). From this 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 the *Caudovirales*. Additionally, modules 1 and 2 share 7 other genes, including a primase, a nuclease of the Cas4 superfamily, and an RNA-primed DNA polymerase of the B family. Modules 2 and 3 are also connected by a tyrosine recombinase that is present, although less commonly, in some plasmids and proviruses outside the *Caudovirales*.

Module 4, *Ligamenvirales*. All members of the order *Ligamenvirales* (families *Rudiviridae* and *Lipothrixviridae*) are grouped into module 4. Three signature genes were detected for this module: the unique, four-helix-bundle MCP, an *S*-adenosylmethionine (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).

Module 5, *Ampullaviridae*. Module 5 encompasses ampullaviruses and family 1 casposons, a recently discovered class of self-synthesizing DNA transposons which employ the Cas1 endonuclease for integration (59–61). The ampullaviruses and the casposons comprise two distinct submodules within this module; the two submodules cluster together in 55% of replicas (Fig. 2A). The submodules are kept together by the protein-primed DNA polymerase B (pPolB), which in the context of archaeal viruses is exclusive to this module as well as the halophilic viruses His1 and His2 (62). The ampullavirus submodule contains 26 shared genes, most of which are refractory to functional annotation. As mentioned above, a glycosyltransferase connects ampullaviruses with members of the *Ligamenvirales*. In addition, a

TABLE 1 Modules in the archaeal virus network

Module	No. of genomes	Robust. genomes[a]	Distinct. genomes[b]	No. of genes	Robust. genes[a]	Distinct. genes[b]	Density[c]	Composition (genome[s])	Composition (gene family[ies])
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*	MCPs 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 plasmids	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 members of the *Fusselloviridae*, *Guttaviridae* and related plasmids	RHH domain protein, DnaA-like AAA^+ ATPase, MCP from fuselloviruses (only *Fusselloviridae*)
9	10	1.00	1.00	27	1.00	0.98	0.47	*Bicaudaviridae* and related "monocaudaviruses"	MoxR-like ATPase, putative integrase, MCP from *Bicaudaviridae*
10	11	0.98	0.88	5	0.98	0.72	0.68	*Pleolipoviridae* and related plasmids	AAA^+ ATPase, major spike protein, integral membrane protein (except for His2), uncharacterized protein

[a] The robustness of a module is the average fraction of replicas in which pairs of members of that module are grouped together.
[b] The distinctiveness of a module is the average fraction of replicas in which members of that module are grouped only with members of the same module. A low value of distinctiveness for a module is indicative that it belongs to a larger supermodule.
[c] The density is the fraction of connections relative to all possible gene-genome pairs in a module. Low density 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, 5, and 6).

detailed analysis of the sequences of the functionally uncharacterized core proteins led to the identification of two DNA-binding proteins, one containing a winged helix-turn-helix (wHTH) DNA-binding domain and the other containing an RHH domain (see Table S2 in the supplemental material). The latter protein provides connections to other archaeal virus modules (Fig. 3).

Module 6, *Turriviridae* and *Sphaerolipoviridae*. A close analysis of module 6 reveals a substructure with three submodules that cluster together in 70 to 90% of replicates. These submodules consist of (i) the *Sphaerolipoviridae*, (ii) the *Turriviridae* and related plasmids/proviruses, and (iii) two plasmids related to betasphaerolipovirus SNJ1 (*Halorubrum saccharovorum* plasmid pZMX101 and *Methanosarcina acetivorans* plasmid pC2A). With the exception of the latter submodule, all other genomes in module 6 contain the A32-like genome packaging ATPase, which is a signature gene of this module in the context of archaeal viruses. Instead, pZMX101 and pC2A join the module by their connection to Natrinema sphaerolipovirus SNJ1 through the rolling-circle replication initiation protein RepA (63). Other remarkable genes from this module are the respective MCPs of sphaerolipoviruses and turriviruses, which are discussed below, because of their similarity with the capsid proteins of some bacterial and eukaryotic viruses.

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

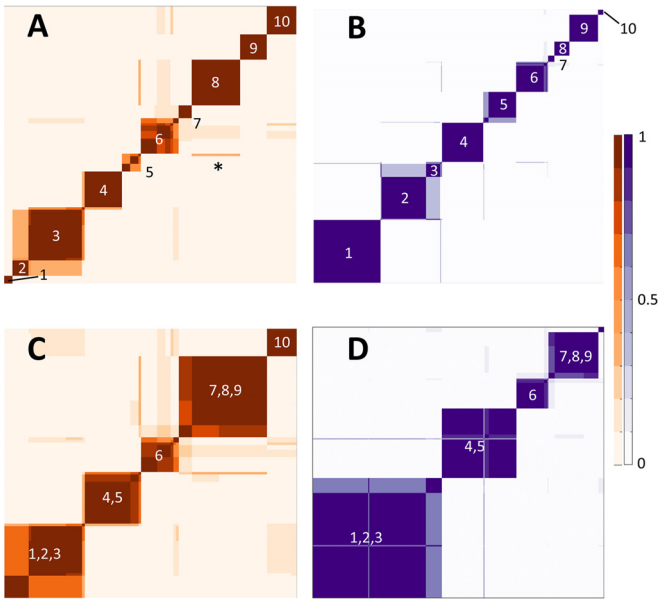

FIG 2 Robustness and cross-similarities of modules (A and B) and supermodules (C and D) in the archaeal virus bipartite network. The module detection algorithm was run in 100 replicas of the original network, yielding 100 alternative partitions of the network. Of these partitions, the one with the highest Barber's modularity index value was selected as the optimal partition; the other 99 were used to assess the robustness of the modules in the optimal partition. (A and C) Heat maps representing the average fraction of replicas in which a pair of genomes was grouped in the same module. Genomes are sorted on both axes based on the module to which they belong in the optimal partition. (B and D) Heat maps for gene families. Dark blocks correspond to robust modules, with the size being proportional to the number of genomes or gene families in the module. Lighter shading within a block suggests the existence of an internal structure, while shaded regions between blocks are indicative of a supermodular structure. The asterisk in panel A denotes the ambiguous assignation of the His1 virus to modules 5 and 8. See the text and Table 1 for a description of the contents of each module.

DnaA-like AAA$^+$ ATPase and an integrase typical of fuselloviruses, but it lacks a detectable homolog of the fusellovirus MCP.

Module 9, *Bicaudaviridae* and related single-tailed viruses. Module 9 encompasses crenarchaeal viruses with large spindle-shaped virions that are decorated with one or two long, tail-like appendages protruding from the pointed virion ends. The module includes Acidianus two-tailed virus, the only classified member of the family *Bicaudaviridae*, as well as 6 unclassified viruses and three viral genomes assembled from metagenomic data (Sulfolobales virus YNP1, Sulfolobales virus YNP2, and hyperthermophilic archaeal virus 2) (see Table S1 in the supplemental material) (27, 40–44). There are four signature genes in this module, including a putative integrase and the bicaudavirus MCP (not detected in metagenomic assemblies), which is unrelated to the capsid proteins of the smaller spindle-shaped viruses from modules 7 and 8 (32). Apart from the connections with fuselloviruses (see above), some members of this module share a MoxR-like ATPase with the haloviruses from module 1. Metagenomic assemblies, including Sulfolobales virus YNP1 and Sulfolobales virus YNP2, harbor two of the signature genes of this module (the putative integrase and a protein with the conserved domain PHA02732, exemplified by open reading frame 52 [ORF52] of Sulfolobus tengchongensis spindle-shaped virus 2 [STSV2]) as well as the RHH domain protein, the DnaA-like AAA$^+$ ATPase, and two uncharacterized proteins present in some other members of the module. Hyperthermophilic archaeal virus 2 (40) is assigned to this module based on a single uncharacterized gene (exemplified by ORF38 of STSV2), which is a signature of bicaudaviruses.

Module 10, *Pleolipoviridae*. Viruses of the family *Pleolipoviridae* have either ssDNA or dsDNA genomes (24) and cluster together in module 10, which encompasses four signature genes. These genes encode both major structural proteins of pleolipoviruses (the spike protein exemplified by protein VP4 of Halorubrum pleomorphic virus 1 [HRPV-1] and the highly divergent integral membrane protein exemplified by VP3 of HRPV-1), an AAA$^+$ ATPase that has been identified as a virion component in

TABLE 2 Connector genes

Family	GI no. of representative sequence	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[a]	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	Archaeoeukaryotic primase	1, 2
19	9628153	Ribonucleotide reductase, large subunit	1, 2

[a] Protein primed in module 5 and RNA primed in modules 1 and 2.

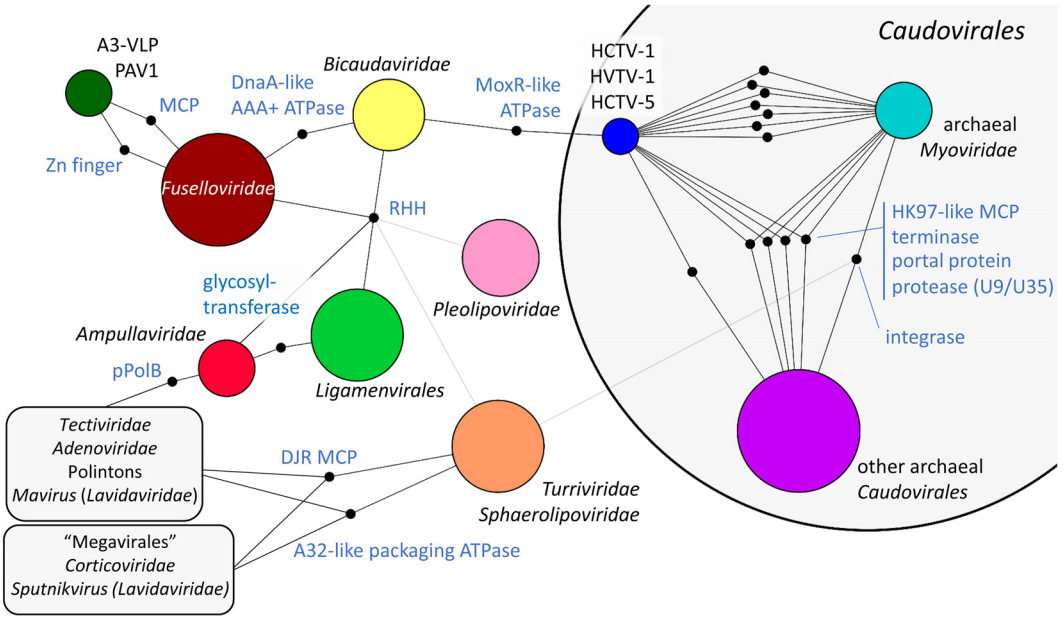

FIG 3 Second-order structure of the archaeal virus network. Large circles represent modules, with their size being proportional to the number of genomes that they encompass. Black dots represent connector genes, i.e., genes whose prevalence in two modules is greater than exp(−1). Light gray edges are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

HRPV-1 (but not in other pleolipoviruses), and an uncharacterized protein (GI:226596535). The His2 virus, the sole member of the genus *Gammapleolipovirus*, encodes a pPolB that connects it to module 5 but is nevertheless unambiguously assigned to the pleolipovirus module on the basis of the rest of its core genes.

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

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

The case of the 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 with the exclusivity of the fusellovirus MCP, suggests that the His1 virus should be assigned to the fusellovirus module. In addition to these genes, His1 shares the DnaA-like AAA$^+$ ATPase with fuselloviruses and bicaudaviruses and shares a glycosyltransferase of the GT-B superfamily with ampullaviruses and members of the *Ligamenvirales*.

The supermodular structure of the network. To explore the existence of a 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. 2C and D): (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 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.

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

There is one important caveat with regard to the relevance of the supermodules: despite the fact that the supermodules seem to be well supported by their robustness, the supermodular structure of the module-connector gene bipartite network as a whole is not significantly different from the structure of a random network ($P = 0.285$ for comparison of the value of Barber's modularity with 200 random networks with the same degree distribution). This lack of statistical significance of the supermodular structure is probably due to the small number of connector, in particular hallmark, genes, which makes it difficult to evaluate whether the modules in the archaeal virus network are actually arranged in supermodules by using topological criteria only. Instead, the relevance of these supermodules has to be assessed based on biological criteria. More specifically, for each pair of modules, it is important to evaluate the legitimacy of merging based on the particular connector genes that they share. For example, there is a substantial body of structural, biochemical, and comparative genomic data suggesting that all members of the *Caudovirales* have emerged from a common ancestor (5, 8, 65, 66), supporting the consolidation of modules 1, 2, and 3 into a single supermodule. In contrast, members of the *Ligamenvirales* and *Ampullaviridae* share neither architectural similarity nor clear commonalities in the mode of genome replication. The only gene that brings the two modules together encodes a glycosyltransferase, which likely mediates certain aspects of virus-host 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 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 the salterprovirus His1. Notably, with the exception of His1, all of these viruses infect hyperthermophilic hosts. Thus, glycosyltransferases might confer an 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 *Caudovirales* 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 members of the *Siphoviridae*. The latter group of genomes (denoted module 9c in reference 47) is itself part of a massive community (module 9 in reference 47) that includes lambdoid phages. This community is characterized by intensive gene exchange and a temperate lifestyle. As occurred with archaeal module 3, this larger community lacks signature genes and instead encompasses 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 archaeal module 2 in the archaeon-only network, but they become connector genes in the complete *Caudovirales* network, as the *Myoviridae* are split into 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 the double-jelly-roll MCP and the single-jelly-roll minor capsid protein with turriviruses (Fig. 3).

Connections between archaeal viruses and nonviral MGEs. Our data set included 14 nonviral MGEs: 11 plasmids and 3 casposons. The automatic module detection approach used here placed these elements into modules together with bona fide viruses, recapitulating previously reported observations based on conventional comparative genomics analyses. The nonviral MGEs 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 into the genomes of members of the *Thaumarchaeota* encode pPolB (67) and are included in module 5 together with ampullaviruses and the 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 the 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 the 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 that 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 two pRN1-related plasmids, pSSVi and pSSVx. These 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 the fusellovirus SSV1 or SSV2, both plasmids are encapsidated into spindle-shaped particles that 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, plasmids pSSVi and pSSVx each share two different genes with fuselloviruses. pSSVi is included in the module via genes encoding the SSV1-like integrase and an RHH domain protein, whereas pSSVx encodes the fusellovirus 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 with the Sulfolobus monocaudavirus 2 (SMV2) from module 9. Notably, the RCRE of SMV2 (GenBank accession number YP_009219263) appears to be inactivated for the following reasons: (i) the conserved motif 2 (HUH, where U is a hydrophobic residue) is changed to YLH; (ii) all homologs of SMV2 RCRE contain two catalytic Tyr residues in motif 3 (YxxxY, where x is any amino acid), a signature of superfamily 1 enzymes (73), whereas SMV2 contains only one of the two tyrosines (YVTKN); and (iii) the gene is not conserved in any of the other bicaudaviruses/monocaudaviruses. Furthermore, the RCRE gene is embedded within the genomic neighborhood including several genes encoding proteins annotated as "conjugative plasmid proteins." Thus, it appears that the 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 was shaped, at least partially, 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 here, might provide a complementary and perhaps more comprehensive account of the deep evolutionary connections within the viral world and can be useful for guiding 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 4 additional families remained disconnected from the rest of the virus network. Several unclassified viruses found a 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 heterogeneity in terms of genomic relatedness and the propensity for gene exchange among different groups of viruses. Some modules harbor numerous signature genes, whereas others have few or none. The latter are typically held together by a dense network of shared genes with a patchy distribution within the module and/or by highly prevalent core genes shared with other modules. Most of the modules are linked via connector genes encoding a small set of widespread proteins, most notably the RHH domain-containing transcription factors and glycosyltransferases, neither of which is a viral hallmark protein. It appears more likely that the two genes have been independently acquired from their hosts by viruses within each module or spread between viruses horizontally. Such a lack of strong connectivity among the modules, with the exception of the *Caudovirales* (modules 1 to 3) and, to a lesser extent, spindle-shaped viruses (modules 7 and 8), indicates that most of the viral groups within the archaeal virosphere are evolutionarily distinct. In stark contrast, 5 of the 10 modules include capsidless MGEs, suggesting that gene flow between bona fide viruses and such elements played a key role in molding the archaeal virosphere. In this respect, the origin and evolution of archaeal viruses mirror those of the eukaryotic virosphere, where connections between viruses and various capsidless elements involve all major groups of viruses and encompass multiple transitions from capsidless elements to bona fide viruses and vice versa (74, 77, 78). Importantly, apart from the modules including the *Caudovirales* and *Turriviridae-Sphaerolipoviridae*, archaeal viruses do not display robust evolutionary connections to eukaryotic or bacterial viruses. This is particularly true for viruses infecting hyperthermophilic crenarchaea, which continue to occupy a unique position within the global virosphere (47).

The observation that nonviral MGEs and archaeal viruses are connected primarily through replication proteins could be of particular significance for understanding the origins of the archaeal virosphere. All viral genomes encompass two major components, namely, determinants of virion formation and those of genome replication. In this context, genome replication modules of some archaeal viruses could be derived from different groups of plasmids and, in the case of protein-primed family B DNA polymerases, from self-synthesizing transposons, the family 1 casposons. In contrast, replication protein genes in other groups of archaeal viruses have been clearly acquired from the host. It has been shown previously that archaeal viruses (and plasmids) from different families that infect taxonomically distant hosts have acquired the genes for replicative minichromosome maintenance (MCM) helicases from their respective hosts on multiple independent occasions (79). Interestingly, archaeal members of the *Caudovirales* from module 1 encode nearly complete archaeon-specific suites of genome replication proteins. For example, HVTV-1 encodes a DNA polymerase, archaeo-eukaryotic primase (AEP), RNase HI, as well as a DNA clamp and its loader (80). Notably, however, some crenarchaeal viruses do not encode any identifiable replication proteins and might therefore employ unique genome replication strategies or encode specific proteins for hijacking the host replication machinery. The origin of the other major component of the viral genomes, namely, determinants of virion

structure, is more difficult to trace. By definition, structural proteins encoded by archaeon-specific viruses have no homologs among bacterial and eukaryotic viruses. Nevertheless, we recently described a case where one of the major nucleocapsid proteins of the tristromavirus Thermoproteus tenax virus 1 (TTV1) has been exapted from the inactivated Cas4-like nuclease (81). Thus, the replication module in many archaeal viruses can be traced to nonviral MGEs or archaeal hosts, whereas structural proteins of archaeal viruses (and viruses in general) can occasionally evolve from cellular proteins that have no *a priori* role in virion formation.

The results of network analyses of archaeal viruses differ from those reported previously for viruses of bacteria and eukaryotes (47) in that the modules of the archaeal virus network, with the exception of the *Caudovirales*, are quite sparsely connected, so much so that although supermodules were identified, their reality could not be supported statistically. Overall, only 9% of the connections in the archaeal network involve members of different modules (excluding the *Caudovirales*), whereas such intermodule connections constitute 25% of the bacterial *Caudovirales* network. The explanation for this sharp distinction is likely to be 2-fold. First, the current sampling of archaeal viruses is likely to be much less representative of their true diversity than the sampling of viruses of bacteria and eukaryotes. Nearly all hyperthermophilic archaea possess clustered regularly interspaced short palindromic repeat (CRISPR)-Cas adaptive immunity loci, often multiple ones (82), which is indicative of the perennial coevolution of these archaea with diverse viromes. However, viruses of archaeal hyperthermophiles outside the *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 nonviral MGEs. 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

This work, including the efforts of Jaime Iranzo and Eugene V. Koonin, was funded by HHS | National Institutes of Health (NIH) (intramural funds of the US Department of Health and Human Services). This work, including the efforts of David Prangishvili, was funded by Agence Nationale de la Recherche (ANR) (program BLANC project EXAVIR).

REFERENCES

1. **Pietilä MK, Demina TA, Atanasova NS, Oksanen HM, Bamford DH.** 2014. Archaeal viruses and bacteriophages: comparisons and contrasts. Trends Microbiol **22:**334–344. http://dx.doi.org/10.1016/j.tim.2014.02.007.
2. **Prangishvili D.** 2013. The wonderful world of archaeal viruses. Annu Rev Microbiol **67:**565–585. http://dx.doi.org/10.1146/annurev-micro-092412-155633.
3. **Prangishvili D, Garrett RA, Koonin EV.** 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. Virus Res **117:**52–67. http://dx.doi.org/10.1016/j.virusres.2006.01.007.
4. **Snyder JC, Bolduc B, Young MJ.** 2015. 40 years of archaeal virology: expanding viral diversity. Virology **479–480:**369–378. http://dx.doi.org/10.1016/j.virol.2015.03.031.
5. **Sencilo A, Jacobs-Sera D, Russell DA, Ko CC, Bowman CA, Atanasova NS, Osterlund E, Oksanen HM, Bamford DH, Hatfull GF, Roine E, Hendrix RW.** 2013. Snapshot of haloarchaeal tailed virus genomes. RNA Biol **10:**803–816. http://dx.doi.org/10.4161/rna.24045.
6. **Atanasova NS, Bamford DH, Oksanen HM.** 2016. Virus-host interplay in high salt environments. Environ Microbiol Rep **8:**431–444. http://dx.doi.org/10.1111/1758-2229.12385.
7. **Pfister P, Wasserfallen A, Stettler R, Leisinger T.** 1998. Molecular analysis of Methanobacterium phage psiM2. Mol Microbiol **30:**233–244. http://dx.doi.org/10.1046/j.1365-2958.1998.01073.x.
8. **Krupovic M, Forterre P, Bamford DH.** 2010. Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol **397:**144–160. http://dx.doi.org/10.1016/j.jmb.2010.01.037.
9. **Krupovic M, Spang A, Gribaldo S, Forterre P, Schleper C.** 2011. A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. Biochem Soc Trans **39:**82–88. http://dx.doi.org/10.1042/BST0390082.
10. **Luo Y, Pfister P, Leisinger T, Wasserfallen A.** 2001. The genome of archaeal prophage PsiM100 encodes the lytic enzyme responsible for autolysis of Methanothermobacter wolfeii. J Bacteriol **183:**5788–5792. http://dx.doi.org/10.1128/JB.183.19.5788-5792.2001.
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. http://dx.doi.org/10.1007/s00705-013-1970-6.
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. http://dx.doi.org/10.1128/JVI.79.14.9097-9107.2005.
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. http://dx.doi.org/10.1155/2013/456318.
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. http://dx.doi.org/10.1128/JVI.00869-09.
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. http://dx.doi.org/10.1016/j.virol.2012.05.036.
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. http://dx.doi.org/10.1016/j.str.2015.07.015.
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. http://dx.doi.org/10.1016/j.str.2013.02.026.
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. http://dx.doi.org/10.1073/pnas.1300601110.
19. **Krupovic M, Bamford DH.** 2008. Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol **6:**941–948. http://dx.doi.org/10.1038/nrmicro2033.
20. **Krupovic M, Koonin EV.** 2015. Polintons: a hotbed of eukaryotic virus, transposon and plasmid evolution. Nat Rev Microbiol **13:**105–115. http://dx.doi.org/10.1038/nrmicro3389.
21. **Krupovic M, Bamford DH.** 2008. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. Virology **375:**292–300. http://dx.doi.org/10.1016/j.virol.2008.01.043.
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. http://dx.doi.org/10.1111/1462-2920.12235.
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. http://dx.doi.org/10.4056/sigs.2645906.
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. http://dx.doi.org/10.1007/s00705-015-2613-x.
25. **Roine E, Kukkaro P, Paulin L, Laurinavicius S, Domanska A, Somer-**

harju P, Bamford DH. 2010. New, closely related haloarchaeal viral elements with different nucleic acid types. J Virol **84:**3682–3689. http://dx.doi.org/10.1128/JVI.01879-09.
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. http://dx.doi.org/10.1016/j.virol.2007.03.005.
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. http://dx.doi.org/10.1111/1462-2920.13079.
28. **Prangishvili D, Krupovic M.** 2012. A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales." Arch Virol **157:**791–795. http://dx.doi.org/10.1007/s00705-012-1229-7.
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. http://dx.doi.org/10.1073/pnas.1518929113.
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. http://dx.doi.org/10.1016/j.virol.2010.03.046.
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. http://dx.doi.org/10.1073/pnas.1203668109.
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. http://dx.doi.org/10.1128/JVI.02941-13.
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. http://dx.doi.org/10.1128/JB.01896-06.
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. http://dx.doi.org/10.1111/j.1462-2920.2011.02662.x.
35. **Iverson E, Stedman K.** 2012. A genetic study of SSV1, the prototypical fusellovirus. Front Microbiol **3:**200. http://dx.doi.org/10.3389/fmicb.2012.00200.
36. **Prangishvili D, Stedman K, Zillig W.** 2001. Viruses of the extremely thermophilic archaeon Sulfolobus. Trends Microbiol **9:**39–43. http://dx.doi.org/10.1016/S0966-842X(00)01910-7.
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. http://dx.doi.org/10.1111/j.1462-2920.2009.02009.x.
38. **Contursi P, Fusco S, Cannio R, She Q.** 2014. Molecular biology of fuselloviruses and their satellites. Extremophiles **18:**473–489. http://dx.doi.org/10.1007/s00792-014-0634-0.
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. http://dx.doi.org/10.1038/4361101a.
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. http://dx.doi.org/10.1111/j.1462-2920.2010.02266.x.
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. http://dx.doi.org/10.1007/s00792-013-0591-z.
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. http://dx.doi.org/10.1128/JVI.79.14.8677-8686.2005.
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. http://dx.doi.org/10.1111/mmi.12503.
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. http://dx.doi.org/10.1128/JVI.03098-15.
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. http://dx.doi.org/10.1128/JVI.07196-11.
46. **Held NL, Whitaker RJ.** 2009. Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. Environ Microbiol **11:**457–466. http://dx.doi.org/10.1111/j.1462-2920.2008.01784.x.
47. **Iranzo J, Krupovic M, Koonin EV.** 2016. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. mBio **7:**e00978-16. http://dx.doi.org/10.1128/mBio.00978-16.
48. **Li W, Godzik A.** 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:**1658–1659. http://dx.doi.org/10.1093/bioinformatics/btl158.
49. **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:**3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.
50. **Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF.** 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res **29:**2994–3005. http://dx.doi.org/10.1093/nar/29.14.2994.
51. **Rosvall M, Bergstrom CT.** 2008. Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci U S A **105:**1118–1123. http://dx.doi.org/10.1073/pnas.0706851105.
52. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
53. **Meier A, Söding J.** 2015. Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. PLoS Comput Biol **11:**e1004343. http://dx.doi.org/10.1371/journal.pcbi.1004343.
54. **Söding J.** 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics **21:**951–960. http://dx.doi.org/10.1093/bioinformatics/bti125.
55. **Makarova KS, Wolf YI, Koonin EV.** 2015. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between Thermococcales, Methanococcales, and Methanobacteriales. Life (Basel) **5:**818–840. http://dx.doi.org/10.3390/life5010818.
56. **Barber MJ.** 2007. Modularity and community detection in bipartite networks. Phys Rev E Stat Nonlin Soft Matter Phys **76:**066102. http://dx.doi.org/10.1103/PhysRevE.76.066102.
57. **Marquitti FM, Guimaraes PR, Pires MM, Bittencourt LF.** 2014. Modular: software for the autonomous computation of modularity in large network sets. Ecography **37:**221–224. http://dx.doi.org/10.1111/j.1600-0587.2013.00506.x.
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. http://dx.doi.org/10.1073/pnas.1633576100.
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. http://dx.doi.org/10.1016/j.mib.2016.01.006.
60. **Béguin P, Charpin N, Koonin EV, Forterre P, Krupovic M.** Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. Nucleic Acids Res, in press. http://dx.doi.org/10.1093/nar/gkw821.
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. http://dx.doi.org/10.1093/nar/gkv1180.
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. http://dx.doi.org/10.1016/j.virol.2006.02.005.
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. http://dx.doi.org/10.1128/JB.00131-16.
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. http://dx.doi.org/10.1128/JB.05101-11.

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. http://dx.doi.org/10.1016/j.coviro.2011.06.001.
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. http://dx.doi.org/10.1073/pnas.1303047110.
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. http://dx.doi.org/10.1186/1741-7007-12-36.
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. http://dx.doi.org/10.1371/journal.pone.0049044.
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. http://dx.doi.org/10.1046/j.1365-2958.1999.01573.x.
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. http://dx.doi.org/10.1016/j.virol.2007.01.035.
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. http://dx.doi.org/10.1128/JB.182.17.4998-5000.2000.
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. http://dx.doi.org/10.1128/AEM.00525-13.
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. http://dx.doi.org/10.1093/nar/20.13.3279.
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. http://dx.doi.org/10.1128/MMBR.00049-13.
75. **Koonin EV, Senkevich TG, Dolja VV.** 2006. The ancient virus world and evolution of cells. Biol Direct **1:**29. http://dx.doi.org/10.1186/1745-6150-1-29.
76. **King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed).** 2012. Virus taxonomy. Classification and nomenclature of viruses. Ninth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press, San Diego, CA.
77. **Koonin EV, Dolja VV, Krupovic M.** 2015. Origins and evolution of viruses of eukaryotes: the ultimate modularity. Virology **479–480:**2–25. http://dx.doi.org/10.1016/j.virol.2015.02.039.
78. **Krupovic M.** 2013. Networks of evolutionary interactions underlying the polyphyletic origin of ssDNA viruses. Curr Opin Virol **3:**578–586. http://dx.doi.org/10.1016/j.coviro.2013.06.010.
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. http://dx.doi.org/10.1093/molbev/msq161.
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. http://dx.doi.org/10.1093/nar/gkw322.
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. http://dx.doi.org/10.1186/s13062-015-0093-2.
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. http://dx.doi.org/10.1038/nrmicro3569.