RESEARCH PAPER

RNA Biology 10:5, 803–816; May 2013; © 2013 Landes Bioscience

Snapshot of haloarchaeal tailed virus genomes

Ana Senčilo,[1] Deborah Jacobs-Sera,[2] Daniel A. Russell,[2] Ching-Chung Ko,[2] Charles A. Bowman,[2] Nina S. Atanasova,[1] Eija Österlund,[1] Hanna M. Oksanen,[1] Dennis H. Bamford,[1] Graham F. Hatfull,[2] Elina Roine[1,*] and Roger W. Hendrix[2,*]

[1]Department of Biosciences and Institute of Biotechnology; University of Helsinki; Helsinki, Finland; [2]Department of Biological Sciences; University of Pittsburgh; PA USA

Keywords: virus, haloarchaea, Archaea, tailed virus, bacteriophage, genome, sequence

Abbreviations: AdoMet-MTase, S-adenosylmethionine-dependent methyltransferase; Cas, CRISPR associated; CDD, conserved domains database; CRISPR, clustered regularly interspaced short palindromic repeats; dNTP, Deoxyribonucleotide triphosphate; ds, double-stranded; HE, homing endonuclease; HGT, horizontal gene transfer; HJR, holliday junction resolvase; HPR, hyperplastic region; *Hqw*, *Haloquadratum walsbyi*; MCM, minichromosome maintenance; MCP, major capsid protein; MGM, modified growth medium; ORF, open reading frame; PesLS, promoter stem loops; Pham, phamily; RFC, replication factor C; RNR, ribonucleotide reductase; SD, shine dalgarno; SNase, staphylococcal nuclease; SW, salt water; TerL, terminase large subunit; Thy, thymidylate; TMP, (1) tape measure protein; (2) thymidine monophosphate

The complete genome sequences of archaeal tailed viruses are currently highly underrepresented in sequence databases. Here, we report the genomic sequences of 10 new tailed viruses infecting different haloarchaeal hosts. Among these, only two viral genomes are closely related to each other and to previously described haloviruses HF1 and HF2. The approximately 760 kb of new genomic sequences in total shows no matches to CRISPR/Cas spacer sequences in haloarchaeal host genomes. Despite their high divergence, we were able to identify virion structural and assembly genes as well as genes coding for DNA and RNA metabolic functions. Interestingly, we identified many genes and genomic features that are shared with tailed bacteriophages, consistent with the hypothesis that haloarchaeal and bacterial tailed viruses share common ancestry, and that a viral lineage containing archaeal viruses, bacteriophages and eukaryotic viruses predates the division of the three major domains of non-viral life. However, as in tailed viruses in general and in haloarchaeal tailed viruses in particular, there are still a considerable number of predicted genes of unknown function.

Introduction

Archaeal virology is a relatively young research discipline in only its fourth decade. During that time, the field has expanded substantially with over 50 viruses described in detail and many more isolated.[1-5] Most of the archaeal virus isolates originate from regions that are heated to high temperatures geothermally or are hypersaline.[2,6]

To date, approximately 60 isolated viruses infecting halophilic archaea have been described,[7] and 14 of the genomes have been sequenced.[8-18] Seven of them form a recently described group of haloarchaeal pleomorphic viruses, HRPV-1, HRPV-2, HRPV-3, HRPV-6, HHPV-1, HGPV-1 and His2.[6,13,15,16,18,19] These viruses are distinctive due to their unusual morphotypes and diverse genomes despite their general relatedness.[6,15,16,18,19] Other sequenced haloarchaeal viruses are the spindle-shaped virus His1 and two icosahedral viruses SH1 and HHIV-2.[8,17,20] Molecular and structural studies on these viruses have also been described.[12,15,19,21-25]

The majority of haloarchaeal viruses contain double-stranded (ds) DNA genomes and are tailed, with morphological similarity to the tailed bacteriophages.[1,5,26-28] In contrast to the approximately 6,000 reported and more than 650 sequenced tailed phages, there are approximately 50 reported tailed viruses infecting haloarchaea.[7] The complete genomic sequences of four of these (φCh1, HF1, HF2 and BJ1) have been described.[9-11,14] Additional sequence information is available from the recently reported collection of 42 "environmental haloviruses."[29] A few related proviral regions have also been identified in the genomes of several euryarchaeal orders[30,31] as well as in the organisms belonging to *Thaumarchaea* and *Crenarchaea*,[32,33] although no tailed viruses have been isolated for crenarchaeal hosts.

Studies on the genomes of tailed bacteriophages have revealed a tremendous genetic diversity characteristic to these viruses.[34-37] The diversity is created by a variety of mechanisms such as point mutations, insertions (including moron acquisition), deletions and multiple types of recombination.[38] Illegitimate recombination is considered to be a major evolutionary force creating mosaic genomes, consisting of regions derived from a common gene pool.[39-41]

Comparative genomic analyses have shown that as their bacteriophage counterparts, archaeal tailed (pro)viruses have mosaic genomes, likely evolving through similar mechanisms.[9-11,31] Limited studies on archaeal tailed viruses have shown that they

*Correspondence to: Elina Roine, Email: Elina.roine@helsinki.fi; Roger W. Hendrix, Email: rhx@pitt.edu
Submitted: 12/10/12; Revised: 02/18/13; Accepted: 02/19/13
http://dx.doi.org/10.4161/rna.24045

Table 1. Properties of the haloarchaeal viruses and their genomes involved in this study

Virus	Morpho-type	Size (bp)	GC (%)	# ORFs	# tRNAs	Ends[a]	Accession	Origins[b]	Host	Reference	Sequencing information
HCTV-1	sipho	103257	57.0	160	1	739 bp DTR	KC292029	MdS	'Har. californiae'	1	454 and Sanger libraries/5,294 reads/13 × coverage (avg. across the two)
HCTV-2	sipho	54291	68.1	86	0	Circ perm	KC292028	SS	'Har. californiae'	5	Ion Torrent/232,007 reads, no Sanger/495 × coverage
HCTV-5	sipho	102105	57.6	166	1	583 bp DTR	KC292027	SS	'Har. californiae'	5	Sanger only/1,550 reads/37 primers/ length ~760 bp/ 11 × coverage
HGTV-1	myo	143855	50.4	281	36*	Circ perm (nick)	KC292026	SS	*Halogranum* sp SS5–1	5	454 and Sanger libraries/ 3,029 reads/no primer walks/64 × coverage
HHTV-1	sipho	49107	56.5	74	0	Circ perm	KC292025	MdS	*Har. hispanica* (62.5% GC)	1	454 + Sanger walks/13 primers 3,047 reads/26 × coverage
HHTV-2	sipho	52643	66.6	88	0	Circ perm	KC292024	SS	*Har. hispanica* (62.5% GC)	5	454 and Sanger walks/12 primers/5,402 reads/26 × coverage
HRTV-4	sipho	35722	59.5	73	0	Circ perm	KC292023	MdS	*Halorubrum* sp s5a-3	5	454 + Sanger walks/4 primers/3,430 reads/26 × coverage
HRTV-5	myo	76134	56.4	118	4	271 bp DTR	KC292022	MdS	*Halorubrum* sp s5a-3	5	Sanger only 18 primers/916 reads/9 ´ × coverage
HRTV-7	myo	69048	59.6	105	1	340 bp DTR	KC292021	MdS	*Halorubrum* sp B2–2	5	Sanger only/22 primers/995 reads/avg. length 788 bp/11 × coverage
HRTV-8	myo	74519	57.1	124	4	346 bp DTR	KC292020	SS	*Halorubrum* sp B2–2	5	Sanger only/30 primers/950 reads/avg. length 757 bp/9 × coverage

*Two tRNAs contain an intron; [a]DTR, direct terminal repeat; Circ perm, circularly permuted; [b]MdS, Margherita di Savoia; SS, Samut Sakhon.

share a similar genome organization and a set of functionally conserved proteins with tailed bacteriophages, reflecting common themes in virion architecture and morphogenesis.[9,31] Such similarity between viruses infecting hosts from different domains of life can be explained by common ancestry as proposed by the virus lineage model.[42-45] According to this model, the viral universe can be divided into a small number of viral lineages based on structural components, each of which originated prior to the separation of hosts into the three domains of life.[42-45] Archaeal tailed viruses have been proposed to fall into the HK97 lineage together with bacterial tailed viruses and herpes viruses that infect eukaryotic hosts.[42,44,45] Even though the membership of archaeal tailed viruses in this lineage is strongly supported by the predicted similarities in virion assembly and structure, bacterial and archaeal tailed viruses may have much more in common.[31] This issue has been rather poorly investigated due to the lack of complete genome sequences of archaeal tailed viruses.

In an effort to isolate and describe more archaeal viruses, recent studies have concentrated on high-salt environments where haloarchaea dominate.[1,5] A total of 28 new tailed viruses infecting haloarchaea were isolated.[1,5] Here, we report the genome sequences of 10 viruses from this collection.

Results

The 10 haloarchaeal tailed viruses described here include six siphoviruses and four myoviruses (**Table 1**). The viruses were isolated from Samuth Sakhon, Thailand and Margherita di Savoia, Italy (**Table 1**), and they infect either well-defined species of haloarchaea (*Haloarcula hispanica*), tentative species

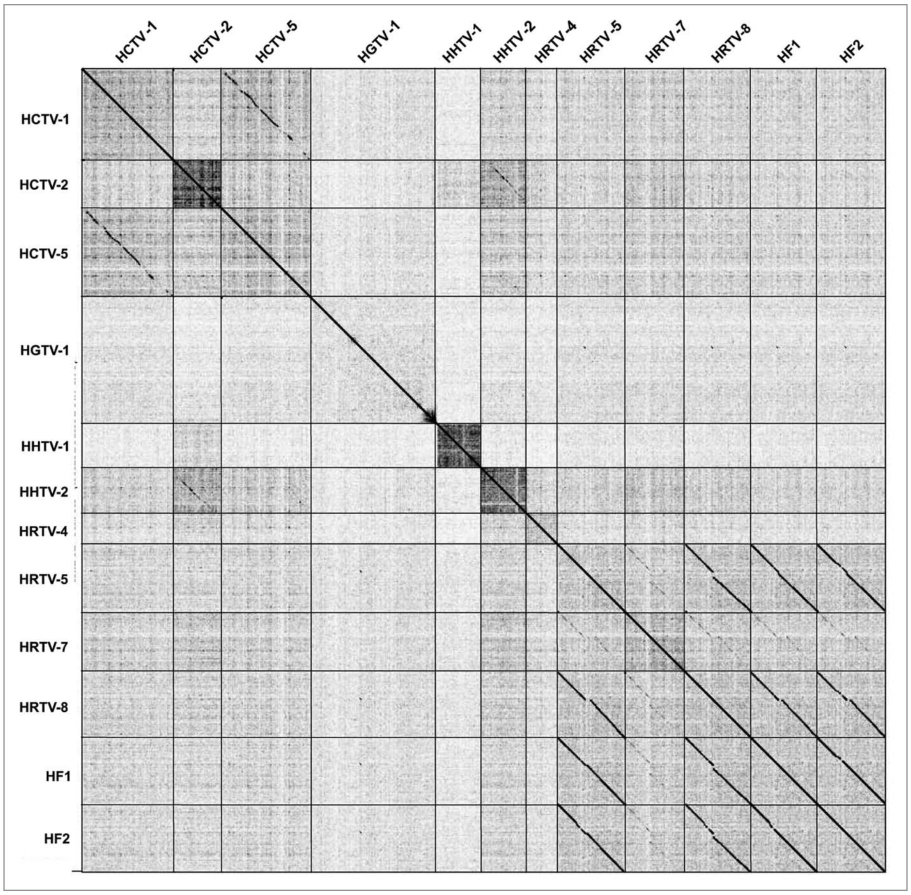


Figure 1. Nucleotide sequence comparison of the studied haloarchaeal tailed viruses and earlier described haloviruses HF1 and HF2[10,11] represented by a dot plot. The dot plot of concatenated viral nucleotide sequences was generated in Gepard.

("Haloarcula californiae") or haloarchaeal isolates assigned to a genus according to the 16S rDNA sequence.[5] Siphoviruses HCTV-1, HCTV-5 and HRTV-4 were shown to infect only the isolation host, whereas for the rest of the viruses, more than one susceptible host was identified.[5] All of the viruses described here form clear plaques and none are obviously temperate, although we note that HCTV-5, HRTV-5, HRTV-7 and HRTV-8 encode putative integrases.

Relationships of haloarchaeal tailed virus genomes. The genomic DNAs were purified and sequenced as described in Materials and Methods. All the viruses contain linear dsDNA genomes ranging in size from 35–144 kb. The ends of the genomes are either circularly permuted (HCTV-2, HHTV-1, HHTV-2, HRTV-4) or they have direct repeats of several hundred base pairs (HCTV-1, HCTV-5, HRTV-5, HRTV-7, HRTV-8); the nature of the HGTV-1 ends is unclear.

We compared the nucleotide sequences of all 10 genomes to each other using Gepard, including the previously described genomic sequences of HF1 and HF2 (**Fig. 1**).[10,11] These genomes are genetically diverse, but several patterns emerge. First, HRTV-5 and HRTV-8 are closely related to each other and form a cluster of related genomes with HF1 and HF2 (**Fig. 1**). HRTV-7 is related to this group with evident regions of sequence similarity, although the strength of the relationship is much weaker (**Fig. 1**). This is reminiscent of the relationships described for the mycobacteriophages, where all five of these genomes would be grouped into a single cluster that would then be subdivided into subclusters according to the strengths of these similarities.[46] Second, HCTV-1 and HCTV-5 show strong sequence similarity and can be considered to be within a common cluster. Third, HCTV-2 and HHTV-2 show weak but easily recognizable nucleotide sequence similarity. Database searches also revealed four predicted proviruses in four euryarchaeal genomes that are related to the viruses described here (**Supplemental Material**).

Haloarchaeal virus protein phamilies. The new genomes were annotated for ORFs, tRNAs and other features (**Figs. 3–5; Figs. S7–9**, **Tables S3–12**). A database was constructed using these 10 newly sequenced genomes as well as the previously published HF1 and HF2 sequences[10,11] using Phamerator.[47] A total of 1,491 putative ORFs was assembled into 966 phamilies (phams) (**Table S13**). Phamily (or pham) is a group of genes the products of which share at least 32.5% identical amino acids (ClustalW) and have BlastP E-value of 10^{-50} or less. Remarkably, 726 of these (75%) are orphams, i.e., phams containing only one member. Most of the phams with more than one member contain ORFs from closely related viruses.

The Phamerator database table provides a simple representation of the genome relationships based on their gene content

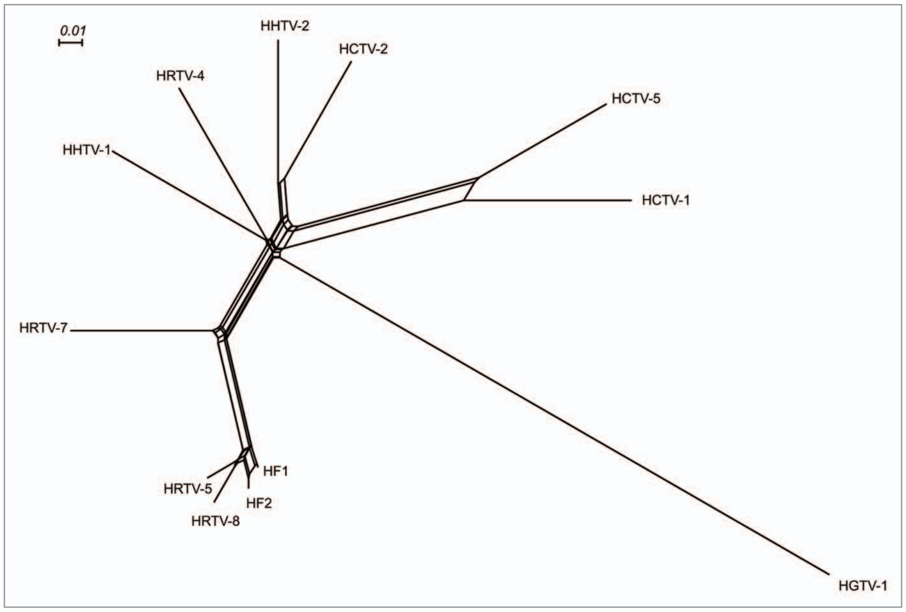


Figure 2. The relationships between the studied haloarchaeal tailed viruses and haloviruses HF1 and HF2[10,11] based on their genome contents. The tree was constructed according to the number of phams shared between the different viruses and depicted using SplitsTree program.[92]

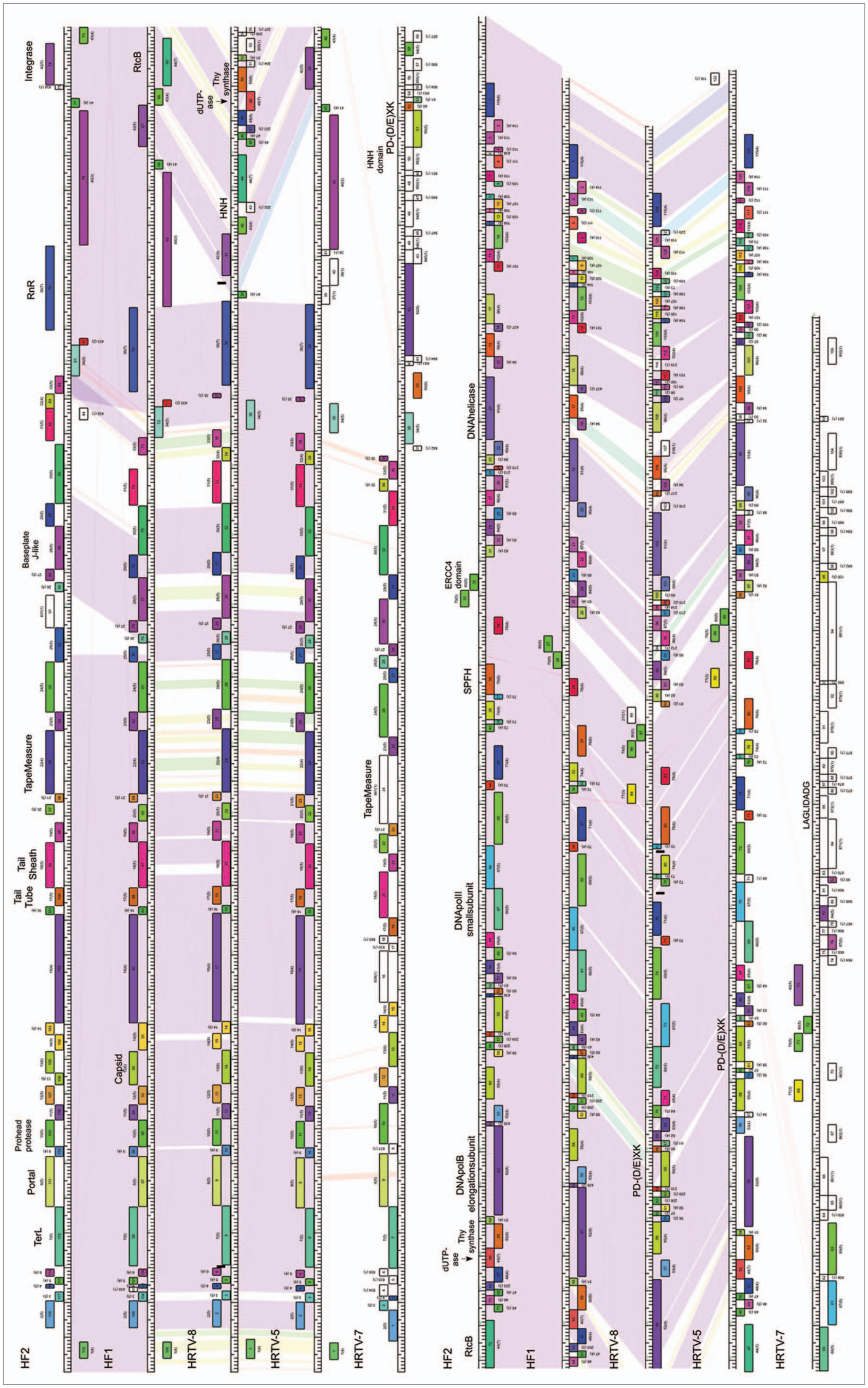
using Splitstree (**Fig. 2**). In general, these relationships reflect those seen in the dotplot nucleotide sequence comparison. For example, HRTV-5 and HRTV-8 clearly group with HF1 and HF2, with HRTV-7 being a more distant relative. HCTV-1 and HCTV-5 group together, as do HHTV-2 and HCTV-2, but in these cases, the relationships are more distant and likely reflect the sharing of only a small number of common genes. Comparative genome maps of these three groups are shown in **Figures 3–5**, respectively, and annotated genome maps of the other viruses are shown in the **Supplemental Material** (**Figs. S7–9**).

Genome duplications and conserved phamilies. Some phams have two or more members from the same virus, reflecting possible genome duplications (**Table S13**). Such examples are the most abundant in HGTV-1 (phams 287, 468, 487, 505, 605; **Fig. S7**). The largest phamily (Pham 14) is composed of nine members, including ORFs from closely related viruses HF1 (ORF91), HF2 (ORFs 103 and 104), HRTV-5 (ORFs 15 and 16), HRTV-7 (ORFs 14 and 15) and HRTV-8 (ORFs 15 and 16). This pham has two members in each of the viruses, except for HF1, where HF1 ORF91 is a fusion of two ORFs found in the other viruses. All these ORFs are located immediately downstream of the major capsid protein (MCP) genes and encode proteins of unknown function. HF2 ORF103 was annotated as a head-tail adaptor protein.[10] Other two large phams (Phams 50 and 52) include eight members each containing ORFs coding for thymidylate synthase and DNA polymerase elongation subunit, respectively. They are found in the majority of these viruses but not in HCTV-2, HHTV-1, HHTV-2 and HRTV-4. Phams 36, 44 and 49 have seven members each coding for ribonucleotide reductase, RtcB-like protein and trimeric dUTPase. Members of the three phams are in half of the viruses in this study and were not found in HCTV-1, HHTV-1, HHTV-2, HRTV-4 and HRTV-7.

Capsid and DNA packaging protein gene clusters. Genes coding for MCPs were identified in all of the viruses. The MCP genes of HCTV-1, HCTV-5, HRTV-5, HRTV-7 and HRTV-8 were identified using the previously determined N-terminal sequences of related haloarchaeal viruses.[48] For HCTV-2, HHTV-1 and HHTV-2, the gene encoding the MCP was also identified using the data of previously determined N-terminal sequences of related haloarchaeal viruses (Pietilä MK and Bamford DH, personal communication). The MCPs of HGTV-1 and HRTV-4 were predicted to belong to the HK97 MCP family due to the presence of conserved residues recognized by BLAST against the CDD. Homology of the majority of the MCPs across all 10 viruses is not detectible at the amino acid level (less than 20% identity). There are several exceptions including MCPs of the two pairs of the most closely related viruses (HRTV-5, HRTV-8 and HCTV-1, HCTV-5). Each pair of the MCPs shares on average 90% amino acid identity. HRTV-5 and HRTV-8 MCPs are also about 40% identical to that of HRTV-7. HHTV-2 and HCTV-2 MCPs have approximately 70% identical amino acid residues.

Other capsid-associated proteins encoded in these viruses are prohead proteases and Mu gpF-like minor capsid proteins (**Supplemental Material**). Mu gpF homologs are either encoded by their own gene or by a gene fusion, the other part of which codes for a putative portal protein. Such an arrangement has been shown for some of the bacteriophages, but not archaeal tailed

Figure 3 (See opposite page). Genome maps of HRTV-5, HRTV-7, HRTV-8, HF1 and HF2. Nucleotide sequence similarity is depicted by shading between the pairs of the genomes according to color spectrum with purple representing the highest identity. Pham and the number of the pham members are outlined above each ORF. TerL, terminase large subunit; RnR, ribonucleotide reductase; Thy, thymidylate.


HF2
HF1
HRTV-8
HRTV-5
HRTV-7
Integrase
RtcB
RnR
Baseplate J-like
TapeMeasure
Tail Tube
Tail Sheath
Prohead protease
Portal
TerL
Capsid
dUTP-ase
Thy synthase
HNH
HNH domain
PD-(D/E)XK
TapeMeasure
HF2
RtcB
HF1
HRTV-8
HRTV-5
HRTV-7
dUTP-ase
Thy synthase
DNApolB elongationsubunit
DNApolII smallsubunit
SPFH
ERCC4 domain
DNAhelicase
PD-(D/E)XK
PD-(D/E)XK
LAGLIDADG

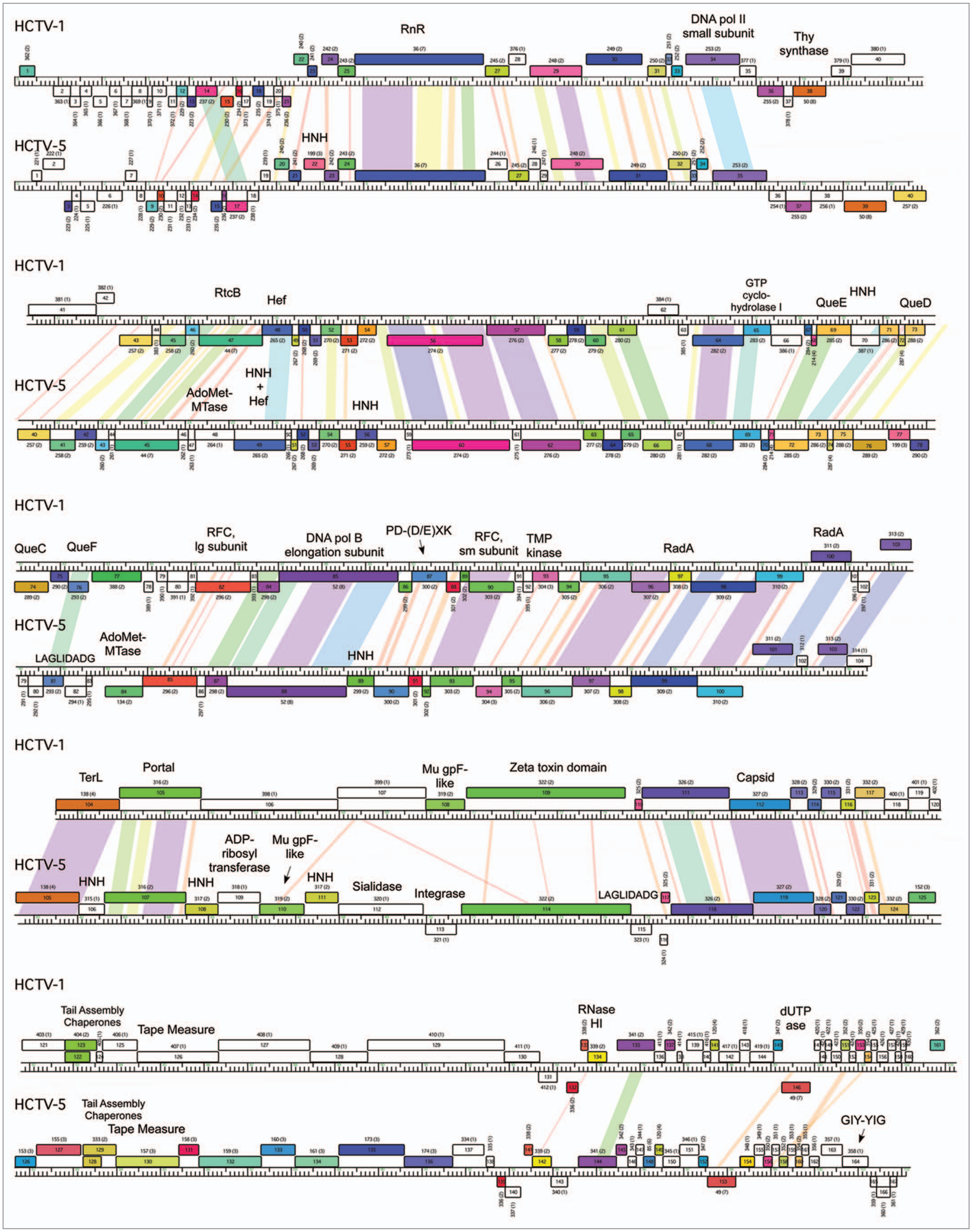

HCTV-1
RnR
DNA pol II small subunit
Thy synthase
HCTV-5
HNH
HCTV-1
RtcB
Hef
GTP cyclo-hydrolase I
QueE
HNH
QueD
HCTV-5
AdoMet-MTase
HNH + Hef
HNH
HCTV-1
QueC
QueF
RFC, Ig subunit
DNA pol B elongation subunit
PD-(D/E)XK
RFC, sm subunit
TMP kinase
RadA
RadA
HCTV-5
LAGLIDADG
AdoMet-MTase
HNH
HCTV-1
TerL
Portal
Mu gpF-like
Zeta toxin domain
Capsid
HCTV-5
HNH
HNH
ADP-ribosyl transferase
Mu gpF-like
HNH
Sialidase
Integrase
LAGLIDADG
HCTV-1
Tail Assembly Chaperones
Tape Measure
RNase HI
dUTPase
HCTV-5
Tail Assembly Chaperones
Tape Measure
GIY-YIG

Figure 4 (See previous page). Genome maps of HCTV-1 and HCTV-5. Nucleotide sequence similarity is depicted by shading as in **Figure 3**. Pham and the number of the pham members are outlined above each ORF. RnR, ribonucleotide reductase; Thy, thymidylate; AdoMet-MTase, S-adenosylmethionine-dependent methyltransferase; RFC, replication factor C; TMP, thymidine monophosphate; TerL, terminase large subunit.

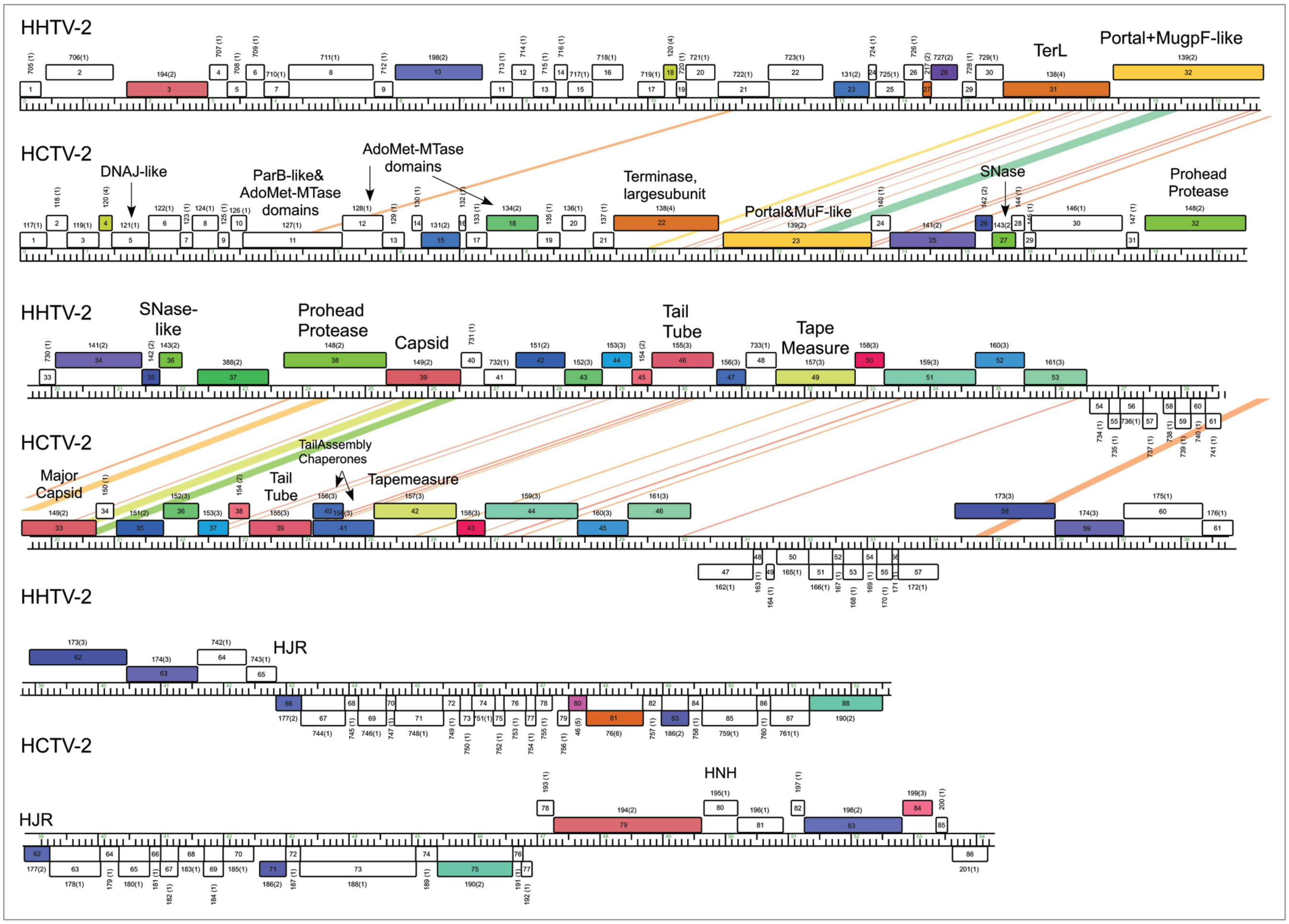


Figure 5. Genome maps of HCTV-2 and HHTV-2. Nucleotide sequence similarity is depicted by shading as in **Figure 3**. Pham and the number of the pham members are outlined above each ORF. TerL, terminase large subunit; AdoMet-MTase, S-adenosylmethionine-dependent methyltransferase; SNase, Staphylococcal nuclease; HJR, Holliday junction resolvase.

viruses.[31] All of the viruses contain a predicted large terminase subunit, although the sequences are diverse, reflecting the diversity of these in the dsDNA tailed bacteriophages. The terminase large subunit gene is typically situated close to one genome end in bacteriophages as also seen in HRTV-8 and the related viruses (**Fig. 3**). In the genomes of HCTV-1 and HCTV-5 that have direct terminal repeats (**Table 1**) the large terminase subunit genes are atypically located near the center of the genome (**Fig. 4**).

Whereas in some virus genomes, the capsid and DNA packaging protein genes are compact and similarly ordered (HGTV-1, HRTV-5, HRTV-7 and HRTV-8), in others the synteny is interrupted by additional ORFs. For example, in HHTV-2 and HCTV-2, a staphylococcal nuclease homolog is encoded between the portal-Mu gpF fusion protein gene and the MCP gene (**Fig. 5**). At a corresponding position, HCTV-1 and HCTV-5 viruses encode a putative protein with Zeta toxin-like domain (**Fig. 4**, **Supplemental Material**).

In addition to the above mentioned putative genes, HCTV-5 is predicted to encode a protein with ADP-ribosyltransferase VIP2 domain and a sialidase. ADP-ribosyltransferase VIP2 domain has been found in several proviruses as a C-terminal part of a fusion protein with Mu gpF in the N-terminal part.[31,49] In the case of HCTV-5 putative VIP2 domain is encoded in a separate ORF just upstream of the putative Mu gpF gene. Downstream of the ORF coding for Mu gpF there is another ORF coding for a putative sialidase. Several genes in this HCTV-5 genome region are flanked by the genes encoding putative homing endonucleases (HEs, see **Fig. 4** and homing endonucleases section).

The organization of the genome regions coding for proteins comprising the phage capsid (minor capsid proteins, prohead protease, MCP) or those that are part of DNA packaging machine (terminase and portal proteins) is generally the same as reported for other archaeal and bacterial tailed viruses.[31,34,37,50]

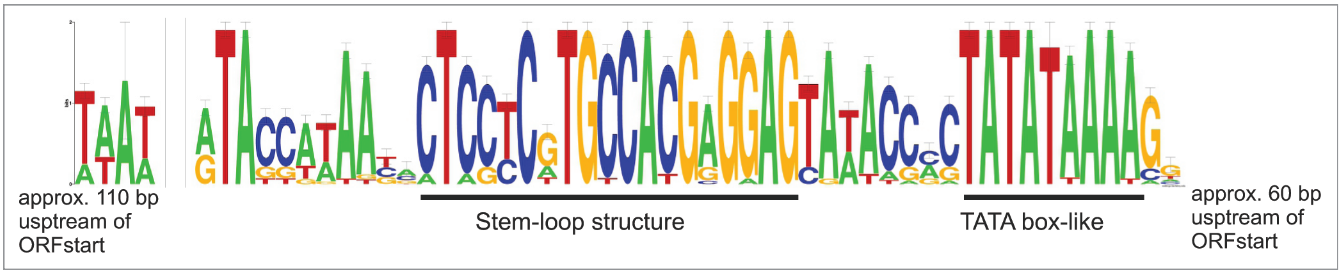


Figure 6. Conserved nucleotides of aligned HGTV-1 repeats. The repeats are found approximately 60 bp upstream of some HGTV-1 genes. Exact positions of the repeats in HGTV-1 genome and their alignment can be found in Supplemental Material (**Fig. S10**). WebLogo was used to generate a graph out of the repeats alignment. The conservation of the nucleotides at individual positions is measured in bits. The TATA box-like and inverted repeat sequences found within the HGTV-1 repeats are underlined and indicated below the sequence logo.

Cluster of genes coding for tail structural and assembly proteins. The length of the tail in tailed viruses is governed by the length of the tail tape measure protein (TMP).[51] Putative ORFs coding for TMPs have been identified in all virus genomes of this study. Predictions were based on PSI-BLAST hits combined with the predicted protein secondary structures, all of which were highly α-helical, as in bacterial viruses.[52] Repeats with regularly spaced aromatic amino acids are characteristic for many of the TMPs.[53] Such eight amino acid long repeats were found in HCTV-1 TMP.

In tailed bacteriophages, the two genes encoding tail assembly chaperones—typically located immediately upstream of the TMP gene—are expressed via a programmed translational frameshift, and this is one of the most highly conserved features of these phages.[54] We have identified several similar examples in the haloarchaeal viruses described here (**Table S2**).

Many myoviruses encode a baseplate J-like protein downstream of the TMP gene. The baseplate J-like protein was shown to be located on the side of the baseplate in P2 virus particles.[55] In addition to a baseplate J-like protein, HGTV-1 also encodes homologs of phage late control D-like protein and baseplate V-like protein, which was shown to be responsible for baseplate assembly and forming the tail spike in P2 phage (**Fig. S2**).[55,56]

DNA metabolism genes. A number of proteins predicted to be involved in deoxyribonucleotide synthesis, DNA replication, recombination and repair, were predicted to be encoded by these viruses. Putative ORFs coding for four key enzymes of deoxyribonucleotide synthesis [ribonucleotide reductase (RNR), thymidylate synthase, thymidylate kinase and dUTPase] were found in HCTV-1, HCTV-5 and HGTV-1 genomes. Ribonucleotide reductases are main players in regulation of deoxyribonucleotide triphosphate (dNTP) levels. In mycobacteriophages, RNR was shown to be important during lytic phage life cycle since it has a role in replacing host enzyme activity.[57]

Siphoviruses HCTV-1 and HCTV-5 encode a substantial number of proteins predicted to play a role in replication. Both viruses encode small and large subunits of replication factor C, DNA polymerase B elongation subunit, DNA polymerase II small subunit and RNase HI (**Fig. 4**). Siphoviruses with smaller genomes (HHTV-1, HHTV-2, HRTV-4 and HCTV-2) encode comparatively small amounts of recognizable proteins involved in DNA and RNA metabolism, albeit some of the encoded proteins may not be found in other viruses from our studied set. For instance, HHTV-1 virus encodes a putative DNA polymerase sliding clamp (**Fig. S8**). In addition to some of the above-mentioned proteins, a virus with the largest genome in our studied set, HGTV-1, encodes a putative DNA ligase and thioredoxin (**Fig. S7**).

HGTV-1 encodes some proteins which may be involved in dsDNA break repair by homologous recombination: putative RecB-like protein, RadA-like protein and Holliday junction resolvase (**Fig. S7**).[58] Some of these components are also found in other viruses from our collection.

HCTV-2 codes for a putative protein with RecJ-like exonuclease S1 domain, which may be implicated in mismatch repair and recombination (**Fig. 5**). HGTV-1 encodes two copies of putative DNA mismatch repair enzymes belonging to endonuclease III family. Moreover, HGTV-1 is predicted to encode nucleotidyltransferase domain of family X DNA polymerases, which takes part both in base excision repair and dsDNA break repair pathways (**Fig. S7**).[59]

tRNAs and RNA metabolism. Some of our viruses, predominantly myoviruses, encode tRNAs. The largest number of tRNA genes is found in HGTV-1 genome. HGTV-1 has 36 tRNA genes, two of which contain introns. HGTV-1 encodes tRNAs for all universal genetic code amino acids. For some of the amino acids there are several tRNA genes with alternative anticodons. We have analyzed codon usage of genes coding for the MCP and tail sheath proteins, which are produced at high numbers during the virus life cycle (data not shown). For both of the genes, HGTV-1 had tRNA genes with anticodons corresponding to overrepresented codons of aspartate, glutamate, phenylalanine and tyrosine. Roughly taken, the HGTV-1 tRNA genes are clustered together and are located between the putative genes coding for DNA polymerase B elongation subunit and a block of putative genes coding for the enzymes involved in deoxyribonucleotide synthesis.

Some of the viruses in our study are predicted to encode enzymes involved in tRNA metabolism. Five of the viruses (HCTV-1, HCTV-5, HRTV-5, HRTV-8 and HGTV-1) encode putative RtcB-like proteins, which have been shown to act as tRNA ligases in *E. coli*.[60] RtcB-like proteins have been found in other prokaryotic viruses and have been suggested to protect viral tRNAs from cleavage by host enzymes.[37] In addition to RtcB-like protein, HGTV-1 encodes RNA ligase and lysyl-tRNA synthetase, which may also be involved in tRNA restriction-repair dynamics in virus-host systems.[61]

HCTV-1 and HCTV-5 viruses are predicted to have a block of genes coding for the enzymes involved in a biosynthesis of a tRNA-specific ribonucleotide, queuosine (**Fig. 4**).[62] Surprisingly, the putative enzymes are those involved in bacterial queuosine, but not archaeosine biosynthesis (**Supplemental Material**). Upstream of the QueCDEF cluster both HCTV-1 and HCTV-5 encode a putative GTP cyclohydrolase I. This enzyme, among other functions, has been shown to be involved in the first steps of queuosine and archaeosine biosynthesis.[62]

Three putative hammerhead ribozymes have been identified in the genomes of HRTV-5 and HRTV-8 (**Fig. S3**). Two of the identified ribozymes are upstream of the genes coding for putative Rad3-like helicases in HRTV-5 and HRTV-8 genomes (at 65,998–66,064 nt and 63,409–63,474 nt positions, respectively) (**Fig. S3**). Earlier, similar hammerhead ribozymes have been identified at equivalent positions in related haloviruses HF1 and HF2 (at 65,540–65,607 nt and at 67,312–67,379 nt, respectively).[63] The other putative ribozyme in HRTV-8 genome is located in intergenic region between putative ORFs 107 and 108 (at 65,506–65,572 nt) (**Fig. S3**). The homologs of the ORF108 are found in HF1, HF2 and HRTV-5 genomes. However, none of the genomes have homologs of HRTV-8 ORF107.

Homing endonucleases. All viruses in this study, except for HHTV-1 and HHTV-2, encode free-standing HEs. The largest number of HEs is found in the HCTV-5 genome, which has 14 putative HEs, the majority of which belong to the HNH family. According to amino acid sequence similarity, HCTV-5 HNH HEs can be divided into three groups (**Figs. S4 and S5**). HCTV-5 encodes an interesting fusion protein containing a Hef nuclease region in its N terminus and an HNH endonuclease in its C terminus. A closely related virus, HCTV-1, encodes only the putative Hef nuclease at a corresponding position, suggesting a recent HNH insertion in HCTV-5 or a deletion in HCTV-1. These phages may benefit from encoding HEs by cleavage of heterologous phage DNA during co-infection.[64]

Repeats in virus genomes. The genome sequences of the viruses were analyzed for the presence of repeats. Most of the identified interspersed repeats are located in the intergenic regions (**Table S14**). Perhaps the most dramatic case is found in HGTV-1 genome. HGTV-1 has a cluster of putative ORFs encoded in the right-hand end of the genome (**Fig. S7**). ORFs in this region are relatively short and most of their products do not have significant matches in the non-redundant protein sequence database. There is a 50 nucleotide long repeat approximately 60 bp upstream of nearly every ORF in this region (**Fig. 6**; **Table S14**).

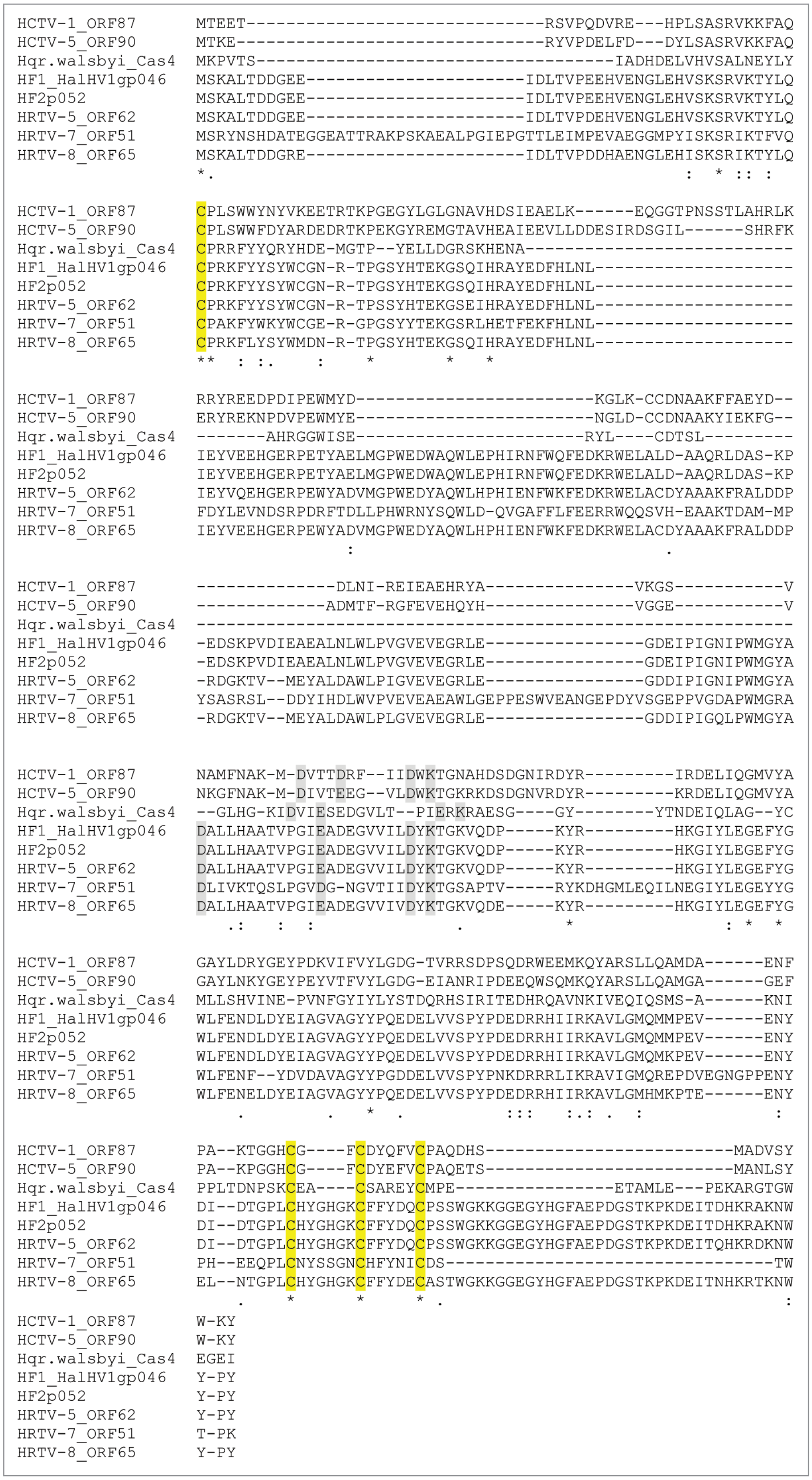
The central part of this repeat contains a 9-base inverted repeat spaced 8 bp from a TATA-box like feature, and likely represents a binding site for a virally encoded transcription factor. However, the TATA-box is positioned approximately 100 bp upstream of the closest ORF, rather than the more typical spacing of approximately 25 bp.[65]

Repeats located in the coding regions of the viral genomes have also been identified. For example, in HCTV-5, ORF114 coding for a protein with Zeta toxin and ParB nuclease motifs contains two types of repeats. One of those is also found in putative ORF coding for ADP-ribosyltransferase (**Table S14**).

In addition to interspersed repeats, a number of tandem repeats ranging from 6–27 nt has been detected in both intergenic and coding regions of the analyzed genomes (**Table S15**). Three different tandem repeats were found in a putative ORF coding for the tape measure protein (TMP) in HCTV-1. TMP are known to evolve by duplications, so the observed repeats could be a trace of this recent evolutionary event in HCTV-1.[66] Tandem repeats seemed to be a non-conserved feature, since even most closely related viruses in our set, HRTV-5 and HRTV-8, showed different patterns of the repeats. In our data set, there was one type of a tandem repeat shared among six viruses. HCTV-2, HCTV-5, HHTV-2, HRTV-4, HRTV-5 and HRTV-8 had a repeat with the consensus sequence "GA(C/G)GA(C/G/A)GA(C/G)GA." In five of the viruses (except for HHTV-2), this repeat was in the coding region of a putative ORF located one or two ORFs upstream of the gene coding for MCP. In these five cases, the sequence codes for a stretch of acidic amino acids glutamate and aspartate.

Search for CRISPR spacer sequences and other related functions. Clustered regularly interspaced short palindromic repeats (CRISPR) are genomic loci, which with the help of CRISPR-associated (Cas) proteins, provide an adaptive immunity against foreign genetic material in prokaryotes.[67-69] Because several of the sequenced haloarchaeal genomes contain predicted CRISPR loci (CRISPRdb), we examined whether any of the viruses described here contain protospacers matching these CIRSPRs. We identified only one possible match, between HRTV-7 and a *Pyrococcus yayanosii* CH1 CRISPR spacer, although with one mismatch and three gaps it may not be functional (**Fig. S6**).

HCTV-1, HCTV-5, HRTV-5, HRTV-7, HRTV-8 and HGTV-1 encode proteins belonging to a large superfamily of RecB-like nucleases (cl00641) that also includes CRISPR/Cas system-associated protein Cas4 (**Table S16**). We have compared these viral proteins with all haloarchaeal proteins belonging to this superfamily from non-redundant protein database available at NCBI (**Table S16**). Viral proteins are rather distantly related to the proteins from the database sharing at most an average of 20% identity at the amino acid level with several hypothetical proteins. The only putative homolog with a predicted function is the *cas4A* gene product of *Haloquadratum walsbyi* (*Hqw*) C23 (YP_005839106.1) sharing 21.1% identical amino acids with ORF90 product of HCTV-5. Despite low similarity, all of the analyzed viral proteins, except for that of HGTV-1, have four conserved cysteine residues and RecB-type nuclease active site residues (**Fig. 7**), which are also found in Cas4 proteins.[70] The RecB type nuclease active site residues of the viral proteins do not align well with the *Hqw* homolog and spacing of the first C-terminal cysteine is not always optimal (three residues apart from the next cysteine). When the HRTV-5 Cas4 homolog was analyzed using Phyre2, the predicted secondary structure



```
HCTV-1_ORF87        MTEET-----------------------------RSVPQDVRE---HPLSASRVKKFAQ
HCTV-5_ORF90        MTKE------------------------------RYVPDELFD---DYLSASRVKKFAQ
Hqr.walsbyi_Cas4    MKPVTS------------------------------------IADHDELVHVSALNEYLY
HF1_HalHV1gp046     MSKALTDDGEE----------------------IDLTVPEEHVENGLEHVSKSRVKTYLQ
HF2p052             MSKALTDDGEE----------------------IDLTVPEEHVENGLEHVSKSRVKTYLQ
HRTV-5_ORF62        MSKALTDDGEE----------------------IDLTVPDEHVENGLEHVSKSRVKTYLQ
HRTV-7_ORF51        MSRYNSHDATEGGEATTRAKPSKAEALPGIEPGTTLEIMPEVAEGGMPYISKSRIKTFVQ
HRTV-8_ORF65        MSKALTDDGRE----------------------IDLTVPDDHAENGLEHISKSRIKTYLQ
                    *.                                               :  * :: :

HCTV-1_ORF87        CPLSWWYNYVKEETRTKPGEGYLGLGNAVHDSIEAELK------EQGGTPNSSTLAHRLK
HCTV-5_ORF90        CPLSWWFDYARDEDRTKPEKGYREMGTAVHEAIEEVLLDDESIRDSGIL------SHRFK
Hqr.walsbyi_Cas4    CPRRFYYQRYHDE-MGTP--YELLDGRSKHENA---------------------------
HF1_HalHV1gp046     CPRKFYYSYWCGN-R-TPGSYHTEKGSQIHRAYEDFHLNL--------------------
HF2p052             CPRKFYYSYWCGN-R-TPGSYHTEKGSQIHRAYEDFHLNL--------------------
HRTV-5_ORF62        CPRKFYYSYWCGN-R-TPSSYHTEKGSEIHRAYEDFHLNL--------------------
HRTV-7_ORF51        CPAKFYWKYWCGE-R-GPGSYYTEKGSRLHETFEKFHLNL--------------------
HRTV-8_ORF65        CPRKFLYSYWMDN-R-TPGSYHTEKGSQIHRAYEDFHLNL--------------------
                    **  : :.    :    *       *   *

HCTV-1_ORF87        RRYREEDPDIPEWMYD------------------------KGLK-CCDNAAKFFAEYD--
HCTV-5_ORF90        ERYREKNPDVPEWMYE------------------------NGLD-CCDNAAKYIEKFG--
Hqr.walsbyi_Cas4    -------AHRGGWISE-----------------------RYL----CDTSL---------
HF1_HalHV1gp046     IEYVEEHGERPETYAELMGPWEDWAQWLEPHIRNFWQFEDKRWELALD-AAQRLDAS-KP
HF2p052             IEYVEEHGERPETYAELMGPWEDWAQWLEPHIRNFWQFEDKRWELALD-AAQRLDAS-KP
HRTV-5_ORF62        IEYVQEHGERPEWYADVMGPWEDYAQWLHPHIENFWKFEDKRWELACDYAAAKFRALDDP
HRTV-7_ORF51        FDYLEVNDSRPDRFTDLLPHWRNYSQWLD-QVGAFFLFEERRWQQSVH-EAAKTDAM-MP
HRTV-8_ORF65        IEYVEEHGERPEWYADVMGPWEDYAQWLHPHIENFWKFEDKRWELACDYAAAKFRALDDP
                                   :                               .

HCTV-1_ORF87        --------------DLNI-REIEAEHRYA---------------VKGS-----------V
HCTV-5_ORF90        -------------ADMTF-RGFEVEHQYH---------------VGGE-----------V
Hqr.walsbyi_Cas4    ------------------------------------------------------------
HF1_HalHV1gp046     -EDSKPVDIEAEALNLWLPVGVEVEGRLE----------------GDEIPIGNIPWMGYA
HF2p052             -EDSKPVDIEAEALNLWLPVGVEVEGRLE----------------GDEIPIGNIPWMGYA
HRTV-5_ORF62        -RDGKTV--MEYALDAWLPIGVEVEGRLE----------------GDDIPIGNIPWMGYA
HRTV-7_ORF51        YSASRSL--DDYIHDLWVPVEVEAEAWLGEPPESWVEANGEPDYVSGEPPVGDAPWMGRA
HRTV-8_ORF65        -RDGKTV--MEYALDAWLPLGVEVEGRLE----------------GDDIPIGQLPWMGYA


HCTV-1_ORF87        NAMFNAK-M-DVTTDRF--IIDWKTGNAHDSDGNIRDYR---------IRDELIQGMVYA
HCTV-5_ORF90        NKGFNAK-M-DIVTEEG--VLDWKTGKRKDSDGNVRDYR---------KRDELIQGMVYA
Hqr.walsbyi_Cas4    --GLHG-KIDVIESEDGVLT--PIERKRAESG----GY--------YTNDEIQLAG--YC
HF1_HalHV1gp046     DALLHAATVPGIEADEGVVILDYKTGKVQDP-----KYR---------HKGIYLEGEFYG
HF2p052             DALLHAATVPGIEADEGVVILDYKTGKVQDP-----KYR---------HKGIYLEGEFYG
HRTV-5_ORF62        DALLHAATVPGIEADEGVVILDYKTGKVQDP-----KYR---------HKGIYLEGEFYG
HRTV-7_ORF51        DLIVKTQSLPGVDG-NGVTIIDYKTGSAPTV-----RYKDHGMLEQILNEGIYLEGEYYG
HRTV-8_ORF65        DALLHAATVPGIEADEGVVIVDYKTGKVQDE-----KYR---------HKGIYLEGEFYG
                       .:   :  :              .          *               : *  *

HCTV-1_ORF87        GAYLDRYGEYPDKVIFVYLGDG-TVRRSDPSQDRWEEMKQYARSLLQAMDA------ENF
HCTV-5_ORF90        GAYLNKYGEYPEYVTFVYLGDG-EIANRIPDEEQWSQMKQYARSLLQAMGA------GEF
Hqr.walsbyi_Cas4    MLLSHVINE-PVNFGYIYLYSTDQRHSIRITEDHRQAVNKIVEQIQSMS-A------KNI
HF1_HalHV1gp046     WLFENDLDYEIAGVAGYYPQEDELVVSPYPDEDRRHIIRKAVLGMQMMPEV------ENY
HF2p052             WLFENDLDYEIAGVAGYYPQEDELVVSPYPDEDRRHIIRKAVLGMQMMPEV------ENY
HRTV-5_ORF62        WLFENDLDYEIAGVAGYYPQEDELVVSPYPDEDRRHIIRKAVLGMQMKPEV------ENY
HRTV-7_ORF51        WLFENF--YDVDAVAGYYPGDDELVVSPYPNKDRRRLIKRAVIGMQREPDVEGNGPPENY
HRTV-8_ORF65        WLFENELDYEIAGVAGYYPQEDELVVSPYPDEDRRHIIRKAVLGMHMKPTE------ENY
                        .        .   *  .          :::   :.: .  :             :

HCTV-1_ORF87        PA--KTGGHCG----FCDYQFVCPAQDHS-------------------------MADVSY
HCTV-5_ORF90        PA--KPGGHCG----FCDYEFVCPAQETS-------------------------MANLSY
Hqr.walsbyi_Cas4    PPLTDNPSKCEA----CSAREYCMPE----------------ETAMLE---PEKARGTGW
HF1_HalHV1gp046     DI--DTGPLCHYGHGKCFFYDQCPSSWGKKGGEGYHGFAEPDGSTKPKDEITDHKRAKNW
HF2p052             DI--DTGPLCHYGHGKCFFYDQCPSSWGKKGGEGYHGFAEPDGSTKPKDEITDHKRAKNW
HRTV-5_ORF62        DI--DTGPLCHYGHGKCFFYDQCPSSWGKKGGEGYHGFAEPDGSTKPKDEITQHKRDKNW
HRTV-7_ORF51        PH--EEQPLCNYSSGNCHFYNICDS-----------------------------------TW
HRTV-8_ORF65        EL--NTGPLCHYGHGKCFFYDECASTWGKKGGEGYHGFAEPDGSTKPKDEITNHKRTKNW
                        .    *      *     * .                                  :

HCTV-1_ORF87        W-KY
HCTV-5_ORF90        W-KY
Hqr.walsbyi_Cas4    EGEI
HF1_HalHV1gp046     Y-PY
HF2p052             Y-PY
HRTV-5_ORF62        Y-PY
HRTV-7_ORF51        T-PK
HRTV-8_ORF65        Y-PY
```

Figure 7. Alignment of the putative Cas4 homologs of haloarchaeal tailed viruses and *Haloquadratum walsbyi* C23. Accession numbers of the proteins are: YP_005839106.1 (*Hqr.walsbyi*_Cas4), NP_861634 (HF1_HalHV1gp046), NP_542546 (HF2p052). Cysteine residues predicted to coordinate FeS center are marked in yellow. Putative nuclease active site residues are marked in grey.

was closest to the AddAB structure (fold library id c3u44B), a helicase-nuclease complex with 97.3% confidence along 48% coverage (data not shown).

Discussion

Here, we report the complete genomic sequences of 10 haloarchaeal tailed viruses, which more than doubles the number of complete haloarchaeal tailed virus genomes previously described. It may seem surprising and unexpected at first blush that some of these viruses, for example HCTV-1 and HCTV-5, have very similar genome sequences and host range yet were isolated from geographically very distant sources (Italy and Thailand in this case). Our small sample of haloarchaeal tailed viruses clearly raises some interesting questions about the ecology and mechanisms of evolution of this group of viruses. Dispersal over large geographic distances and extensive swapping of genes are also seen in the bacteriophages.[71,72] The fascinating questions raised by these observations about the place of viruses in global ecology and the pathways of viral evolution are only beginning to be addressed for those bacteriophages, where there are many hundreds of genome sequences available for analysis. The data we present here for the tailed haloarchaeal viruses, sparse though it is, will lay the groundwork, we hope, for a larger field of study that will develop as more genome sequence, biological and geographic data become available.

Although we found relatedness of some viral genomes in this study, the nucleotide sequence analyses confirm that overall they constitute an underrepresented component of the viral sequence universe because very few similar sequences are present in current databases. Despite the high divergence of our studied viruses, we were able to annotate gene clusters coding for virion structural and assembly proteins, enzymes involved in DNA and RNA metabolism and some additional proteins. Many of these predictions were possible due to similarities between archaeal and bacterial tailed virus proteins. As also shown by previous studies,[31] archaeal tailed viruses studied here seem to share a set of structural proteins with tailed bacteriophages. Shared virion architectural principles play a central role in grouping these distant members of the viral universe into a lineage.[42-45] However, are there any other common characteristics? As in bacteriophages and earlier studied haloarchaeal tailed viruses, the new genomes reported in this study are also mosaic in nature.[10,31,48] Our studies suggest that putative genome shuffling mechanisms such as the promoter stem loops-like (PesLS) repeats in HGTV-1 genome and possibly some features of virus-host interplay (RtcB-like proteins) are shared among bacterial and archaeal tailed viruses. Whether these features were transferred between tailed viruses by horizontal gene transfer (HGT), appeared through convergent evolution or they were vertically inherited from a possible common ancestor remains an open question. Prediction of the inherently bacterial proteins such as bacterial queuosine synthesis enzyme QueF and a protein with Zeta toxin domain encoded in archaeal tailed virus genomes (and previously mycobacteriophages[73]) argues for recent HGT among archaeal viruses and bacteria/bacteriophages. It is clear that all three processes shape genomes of viruses, but the extent and consequence of each requires further investigation.

In search for the genomic repeat sequences, we have found a region in HGTV-1 genome in which most of the genes are flanked by repeats consisting of putative TATA-box and inverted repeats. Besides the putative function in transcription regulation, the repeats could serve as sites for recombination contributing to the genome plasticity of the virus. Similar intergenic repeats have been discovered in T4-like phages.[74] These repeats, the PesLSs, are composed of σ^{70}-like promoter and stem loop structures, which were suggested to be involved in transcription regulation.[74] PesLSs were found mainly in hyperplastic genome regions (HPRs) and were suggested to mediate modular genome shuffling.[74] This idea was supported by the ability of PesLSs to recombine and form mini-circles, which, in turn, were suggested to be encapsidated into virus particles.[74] Similarity of PesLSs to repeats identified in HGTV-1 suggests that it may utilize similar genome shuffling mechanisms as T4-type phages.

Phage T4 is replete with HE genes and encodes 15 HEs in its 168,903-bp long genome.[75] The siphovirus HCTV-5 almost matches this with 14 predicted HE genes representing a greater abundance given that it has a smaller genome than T4. Compared with T4, HCTV-5 also encodes HEs belonging to a broader range of HE families. In bacteriophage T4, the mobility of HEs is dependent on several phage and host-encoded proteins involved in DNA recombination, synthesis and repair.[75] Nevertheless, the presence of these putative proteins in HCTV-5 cannot explain the abundance of HE genes in its genome, because a very close relative HCTV-1 also encodes the same putative proteins involved in DNA recombination, synthesis and repair as HCTV-5, but its genome is almost devoid of the HEs. HCTV-5 is distinct from HCTV-1 in that it encodes a putative integrase between the sialidase and zeta toxin protein genes. However, this integrase is smaller than most phage tyrosine integrases (234 aa) and we can find no nearby sequence corresponding to a putative *attP* site. This integrase may thus be involved in viral DNA rearrangements rather than chromosomal integration.

Collectively, these haloarchaeal viruses encode almost a full set of proteins required for elongation and Okazaki fragment maturation stages of DNA replication. Interestingly, we were not able to identify proteins which would be involved in replication initiation processes, such as ssDNA-binding protein or minichromosome maintenance (MCM) proteins. The only known archaeal tailed virus encoding MCM is BJ1.[14] This contrasts with the studies on archaeal proviruses, in which MCM was the most widespread replication-associated protein.[31]

Our understanding of the CRISPR-Cas systems in prokaryotes, the adaptive defense system against invading genetic material such as viral genomes,[67,68] is growing rapidly. Approximately 90% of sequenced archaeal genomes contain CRISPR-Cas systems,[76] but a recent report on haloarchaeal genomes shows that a complete CRISPR-Cas system can be found in only 12 out of 21 of these sequenced haloarchaeal genomes.[77] Accordingly, in the total of approximately 760 kb of viral sequence, we found only one potential sequence to match a CRISPR spacer in a distant euryarchaeal host (**Fig. S6**). Although as also noted previously,[76-78] the CRISPR-Cas defense system in haloarchaea may not be the primary resistance mechanism against viruses, it is also possible that the diversity of hosts and viruses is greater than previously understood, as suggested by the virus genomes described here; a much larger database of host and viral sequences may thus be necessary to identify such spacer-protospacer matches.

Six new viral genomes as well as HF1 and HF2 contain genes encoding putative PD-(D/E)XK nucleases. Such RecB-type nucleases are also found encoded by the *cas4* genes belonging to CRISPR-Cas gene clusters.[70,79] A recent report by Zhang et al. (2012) shows that Sso0001, a Cas4 homolog of *Sulfolobus solfataricus*, is an iron-sulfur (FeS) cluster protein that forms multimers in a similar way as many other related nucleases. In *Thermoproteus tenax,* Cas4 protein has also been shown to form complexes with Cas1 and Cas2 proteins[80] that are suggested to be involved in the spacer acquisition pathway. Although the virally encoded proteins could play roles in viral replication or recombination, is it possible that they also play roles in immune defense? An alignment of the putative viral and host homologs shows that the four cysteins involved in the formation of FeS-cluster[70] and some of the nuclease active site residues (except for HGTV-1 homolog) can be aligned, although the viral versions of the predicted protein are bigger in size and more variable outside the active site regions (**Fig. 7**). Plausibly, the viral forms of Cas4 could be beneficial for the infection if they could be incorporated into the host complexes and, thus, interfere with the acquisition of the new spacer sequences of the infecting virus. Interestingly, in this study, viruses containing the putative Cas4 homologs infect haloarchaeal species reported to contain complete CRISPR-Cas systems.[77]

There are several lines of evidence that haloarchaeal tailed viruses share many common features with tailed bacteriophages. This is, however, in contrast to the results obtained using the pham analysis, which shows that 75% of the phams are orphams, protein families containing only one member among the studied set of 10 viruses. Not only does it emphasize the underrepresentation of these viruses in the known viral universe, but also the lack of knowledge we have about the genes carried in the archaeal viral genomes and their encoded functions.

Materials and Methods

Archaea, viruses, virus growth and purification. Viruses and host strains used to propagate them are listed in **Table 1**. Host cultures were grown aerobically at 37°C using a modified growth medium (MGM) containing 5 g peptone (Oxoid) and 1 g Bacto yeast extract (Difco Laboratories) per liter.[81] Solid and top-layer media contained 14 and 4 g of Bacto agar (Difco Laboratories) per liter, respectively. A 30% (wt/vol) stock solution of artificial salt water (SW) was added to broth, solid and top-layer media to give final concentrations of 23%, 20% and 18% (wt/vol), respectively. The preparation of the artificial SW is described in the Halohandbook (www.haloarchaea.com/resources/halohandbook/Halohandbook_2009_v7.2mds.pdf).

Virus stocks were prepared by using semi-confluent or confluent top-layer agar plates incubated at 37°C for 2–4 d as previously described.[5] Virus particles were purified from the virus stocks using 18% (wt/vol) SW as a buffer as previously described.[5] In brief, virus particles were precipitated with 10% (wt/vol) polyethylene glycol 6000 and after removal of aggregates, the viruses were purified in a linear 5–20% (wt/vol) sucrose gradient by rate-zonal centrifugation (Sorvall AH629 rotor, 104,000 × *g*, 40–95 min, 20°C) and further purified in a CsCl gradient (mean density of 1.5 g/ml) by equilibrium centrifugation (Sorvall AH629 rotor, 79,000 × *g*, 20 h, 20°C). The purified virus was diluted 2-fold with 18% SW without NaCl and concentrated by differential centrifugation (Sorvall T647.5, 114,000 × *g*, 3 h, at 20°C).

DNA extraction, sequencing and annotation of the viral genomes. Genomic DNA was extracted using phenol/chloroform extraction method followed by ethanol precipitation of DNA. Precipitated DNA was then resuspended in TE Buffer (pH 7.5) and sequenced by the Pittsburgh Bacteriophage Institute using Sanger sequencing, 454 sequencing and Ion Torrent sequencing technologies. Raw reads were assembled using Phred/Phrap/Consed (Sanger method) or 454's GS De Novo Assembler and assemblies were then quality controlled using Consed (454 and Ion Torrent methods). Coverage for each viral genome was dependent on the method, with at least 9 × coverage using Sanger method, 26 × coverage using 454 and greater than 450 × coverage using Ion Torrent. Sanger reads were required to resolve weak areas in the assembly. Sequencing specifics are noted in each GenBank file. Accession numbers are shown in **Table 1**.

Viral genomic sequences were annotated using DNA Master (www.cobamide2.bio.pitt.edu/, version 5.22.5 http://cobamide2.bio.pitt.edu/). The annotations were verified based on coding potential graphs generated by GeneMark.hmm for prokaryotes.[82] The start site positions of the predicted open reading frames (ORFs) were refined based on the proximity to the preceding annotated ORF and the presence of the putative Shine-Dalgarno (SD) sequence (GGA GGT GA).[83] As genes are usually tightly packed in viral genomes, the main criterion for choosing the start position was the shorter distance to the stop position of the immediate upstream ORF. For some genes where there are several possible start sites located at similar distances from the preceding ORF, the regions from 0 to +20 nucleotides upstream the putative start sites were inspected for at least 4–5 nt matches to the SD sequence.

The predicted functions of putative gene products were based on the results of similarity searches (blastp and PSI-BLAST) against the NCBI non-redundant protein database and Conserved Domains Database (CDD). The products of putative ORFs located in the genome regions coding for putative structural proteins were also analyzed using Fold and Function Assignment System server (FFAS03). Some of the putative proteins were analyzed with Homology Detection and Structure Prediction by HMM-HMM comparison tool (HHpred)[84] and Protein Homology/Analogy Recognition Engine V 2.0 (PHYRE2).[85]

Sequence analysis. Nucleotide similarity searches were performed using the BLASTN tool available at the NCBI.[86] Tandem repeats finder[87] and discontiguous megablast similarity searches were used to locate repeated sequences. Alignments were done using T-coffee.[88] Identity of proteins at the amino acid sequence level was determined using the EMBOSS alignment tool Needle available at EMBL-EBI. Conserved DNA and protein motifs were visualized with WebLogo.[89] Virus genome maps and gene assignments into phamilies (phams) were done using Phamerator.[47] Two genes were grouped into the same pham if their predicted amino acid sequences showed at least 32.5% amino acid identity according to ClustalW and a BlastP E-value threshold of 10^{-50}. Putative CRISPR sequences in viral genomes were searched for using CRISPRFinder web tool.[90] Viral genome sequences corresponding to CRISPR spacers were retrieved from CRISPRdb database.[76] Putative RNA elements encoded in the genomes were detected by search against RNA families database (Rfam).[91] Phyre2[85] was used to predict the secondary structure and putative functional homologs for some predicted proteins.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank Sari Korhonen, Terhi Kemppinen and Eija Stenius for technical assistance and University of Helsinki for the support to EU ESFRI Instruct Centre for Virus Production and Purification used in this study. This study was also supported by the Helsinki University three year grant (2010–2012) to E.R., by Academy of Finland grant 1129684 (Centre of Excellence in Virus Research 2006–2011) to D.H.B. and by Academy Professor (Academy of Finland) funding grants 255342 and 256518 to D.H.B. Support was also provided by a grant from the National Institutes of Health to R.W.H. (GM51975). We thank Craig L. Peebles for his help in the analysis of the tRNAs. A.S. and N.S.A. are students of the Viikki Doctoral Programme in Molecular Biosciences.

Supplemental Material

Supplemental material may be found here:
www.landesbioscience.com/journals/rnabiology/article/24045

References

1. Kukkaro P, Bamford DH. Virus-host interactions in environments with a wide range of ionic strengths. Environ Microbiol 2009; 1:71-7; http://dx.doi.org/10.1111/j.1758-2229.2008.00007.x.
2. Pina M, Bize A, Forterre P, Prangishvili D. The archeoviruses. FEMS Microbiol Rev 2011; 35:1035-54; PMID:21569059; http://dx.doi.org/10.1111/j.1574-6976.2011.00280.x.
3. Mochizuki T, Sako Y, Prangishvili D. Provirus induction in hyperthermophilic archaea: characterization of *Aeropyrum pernix* spindle-shaped virus 1 and *Aeropyrum pernix* ovoid virus 1. J Bacteriol 2011; 193:5412-9; PMID:21784945; http://dx.doi.org/10.1128/JB.05101-11.
4. Mochizuki T, Krupovič M, Pehau-Arnaudet G, Sako Y, Forterre P, Prangishvili D. Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. Proc Natl Acad Sci USA 2012; 109:13386-91; PMID:22826255; http://dx.doi.org/10.1073/pnas.1203668109.
5. Atanasova NS, Roine E, Oren A, Bamford DH, Oksanen HM. Global network of specific virus-host interactions in hypersaline environments. Environ Microbiol 2012; 14:426-40; PMID:22003883; http://dx.doi.org/10.1111/j.1462-2920.2011.02603.x.
6. Roine E, Oksanen HM. Viruses from the hypersaline environments: current research an future trends. In: Ventosa A, Oren A, Ma Y, eds. Halophiles and Hypersaline Environments. Heidelberg: Springer, 2011:153-72.
7. Sabet S. Halophilic viruses. In: Vreeland R, ed. Advances in Understanding the Biology of Halophilic Microorganisms. New York: Springer, 2012:81-116.
8. Bath C, Dyall-Smith ML. His1, an archaeal virus of the *Fuselloviridae* family that infects *Haloarcula hispanica.* J Virol 1998; 72:9392-5; PMID:9765495.
9. Klein R, Baranyi U, Rössler N, Greineder B, Scholz H, Witte A. Natrialba magadii virus phiCh1: first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. Mol Microbiol 2002; 45:851-63; PMID:12139629; http://dx.doi.org/10.1046/j.1365-2958.2002.03064.x.
10. Tang SL, Nuttall S, Ngui K, Fisher C, Lopez P, Dyall-Smith M. HF2: a double-stranded DNA tailed haloarchaeal virus with a mosaic genome. Mol Microbiol 2002; 44:283-96; PMID:11967086; http://dx.doi.org/10.1046/j.1365-2958.2002.02890.x.
11. Tang SL, Nuttall S, Dyall-Smith M. Haloviruses HF1 and HF2: evidence for a recent and large recombination event. J Bacteriol 2004; 186:2810-7; PMID:15090523; http://dx.doi.org/10.1128/JB.186.9.2810-2817.2004.
12. Bamford DH, Ravantti JJ, Rönnholm G, Laurinavičius S, Kukkaro P, Dyall-Smith M, et al. Constituents of SH1, a novel lipid-containing virus infecting the halophilic euryarchaeon *Haloarcula hispanica.* J Virol 2005; 79:9097-107; PMID:15994804; http://dx.doi.org/10.1128/JVI.79.14.9097-9107.2005.
13. Bath C, Cukalac T, Porter K, Dyall-Smith ML. His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, *Salterprovirus.* Virology 2006; 350:228-39; PMID:16530800; http://dx.doi.org/10.1016/j.virol.2006.02.005.
14. Pagaling E, Haigh RD, Grant WD, Cowan DA, Jones BE, Ma Y, et al. Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. BMC Genomics 2007; 8:410; PMID:17996081; http://dx.doi.org/10.1186/1471-2164-8-410.
15. Pietilä MK, Roine E, Paulin L, Kalkkinen N, Bamford DH. An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. Mol Microbiol 2009; 72:307-19; PMID:19298373; http://dx.doi.org/10.1111/j.1365-2958.2009.06642.x.
16. Roine E, Kukkaro P, Paulin L, Laurinavičius S, Domanska A, Somerharju P, et al. New, closely related haloarchaeal viral elements with different nucleic Acid types. J Virol 2010; 84:3682-9; PMID:20089654; http://dx.doi.org/10.1128/JVI.01879-09.
17. Jaakkola ST, Penttinen RK, Vilén ST, Jalasvuori M, Rönnholm G, Bamford JK, et al. Closely related archaeal *Haloarcula hispanica* icosahedral viruses HHIV-2 and SH1 have nonhomologous genes encoding host recognition functions. J Virol 2012; 86:4734-42; PMID:22357274; http://dx.doi.org/10.1128/JVI.06666-11.
18. Senčilo A, Paulin L, Kellner S, Helm M, Roine E. Related haloarchaeal pleomorphic viruses contain different genome types. Nucleic Acids Res 2012; 40:5523-34; PMID:22396526; http://dx.doi.org/10.1093/nar/gks215.
19. Pietilä MK, Atanasova NS, Manole V, Liljeroos L, Butcher SJ, Oksanen HM, et al. Virion architecture unifies globally distributed pleolipoviruses infecting halophilic archaea. J Virol 2012; 86:5067-79; PMID:22357279; http://dx.doi.org/10.1128/JVI.06915-11.
20. Porter K, Kukkaro P, Bamford JK, Bath C, Kivelä HM, Dyall-Smith ML, et al. SH1: A novel, spherical halovirus isolated from an Australian hypersaline lake. Virology 2005; 335:22-33; PMID:15823603; http://dx.doi.org/10.1016/j.virol.2005.01.043.
21. Porter K, Dyall-Smith ML. Transfection of haloarchaea by the DNAs of spindle and round haloviruses and the use of transposon mutagenesis to identify non-essential regions. Mol Microbiol 2008; 70:1236-45; PMID:19006816; http://dx.doi.org/10.1111/j.1365-2958.2008.06478.x.
22. Porter K, Russ BE, Yang J, Dyall-Smith ML. The transcription programme of the protein-primed halovirus SH1. Microbiology 2008; 154:3599-608; PMID:18957612; http://dx.doi.org/10.1099/mic.0.2008/019422-0.
23. Kivelä HM, Roine E, Kukkaro P, Laurinavičius S, Somerharju P, Bamford DH. Quantitative dissociation of archaeal virus SH1 reveals distinct capsid proteins and a lipid core. Virology 2006; 356:4-11; PMID:16935317; http://dx.doi.org/10.1016/j.virol.2006.07.027.
24. Jäälinoja HT, Roine E, Laurinmäki P, Kivelä HM, Bamford DH, Butcher SJ. Structure and host-cell interaction of SH1, a membrane-containing, halophilic euryarchaeal virus. Proc Natl Acad Sci USA 2008; 105:8008-13; PMID:18515426; http://dx.doi.org/10.1073/pnas.0801758105.
25. Kandiba L, Aitio O, Helin J, Guan Z, Permi P, Bamford DH, et al. Diversity in prokaryotic glycosylation: an archaeal-derived N-linked glycan contains legionaminic acid. Mol Microbiol 2012; 84:578-93; PMID:22435790; http://dx.doi.org/10.1111/j.1365-2958.2012.08045.x.
26. Torsvik T, Dundas ID. Bacteriophage of *Halobacterium salinarium.* Nature 1974; 248:680-1; PMID:4833269; http://dx.doi.org/10.1038/248680a0.
27. Wais AC, Kon M, MacDonald RE, Stollar BD. Salt-dependent bacteriophage infecting *Halobacterium cutirubrum* and *H. halobium.* Nature 1975; 256:314-5; PMID:1143331; http://dx.doi.org/10.1038/256314a0.
28. Ackermann HW, Prangishvili D. Prokaryote viruses studied by electron microscopy. Arch Virol 2012; 157:1843-9; PMID:22752841; http://dx.doi.org/10.1007/s00705-012-1383-y.
29. Garcia-Heredia I, Martin-Cuadrado AB, Mojica FJ, Santos F, Mira A, Antón J, et al. Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. PLoS One 2012; 7:e33802; PMID:22479446; http://dx.doi.org/10.1371/journal.pone.0033802.
30. Luo Y, Pfister P, Leisinger T, Wasserfallen A. The genome of archaeal prophage PsiM100 encodes the lytic enzyme responsible for autolysis of Methanothermobacter wolfeii. J Bacteriol 2001; 183:5788-92; PMID:11544247; http://dx.doi.org/10.1128/JB.183.19.5788-5792.2001.
31. Krupovič M, Forterre P, Bamford DH. 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 2010; 397:144-60; PMID:20109464; http://dx.doi.org/10.1016/j.jmb.2010.01.037.
32. Krupovič M, Spang A, Gribaldo S, Forterre P, Schleper C. A thaumarchaeal provirus testifies for an ancient association of tailed viruses with archaea. Biochem Soc Trans 2011; 39:82-8; PMID:21265751; http://dx.doi.org/10.1042/BST0390082.
33. Heinemann J, Maaty WS, Gauss GH, Akkaladevi N, Brumfield SK, Rayaprolu V, et al. Fossil record of an archaeal HK97-like provirus. Virology 2011; 417:362-8; PMID:21764098; http://dx.doi.org/10.1016/j.virol.2011.06.019.
34. Krupovič M, Prangishvili D, Hendrix RW, Bamford DH. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev 2011; 75:610-35; PMID:22126996; http://dx.doi.org/10.1128/MMBR.00011-11.
35. Brüssow H, Desiere F. Comparative phage genomics and the evolution of *Siphoviridae*: insights from dairy phages. Mol Microbiol 2001; 39:213-22; PMID:11136444; http://dx.doi.org/10.1046/j.1365-2958.2001.02228.x.
36. Kwan T, Liu J, Dubow M, Gros P, Pelletier J. Comparative genomic analysis of 18 *Pseudomonas aeruginosa* bacteriophages. J Bacteriol 2006; 188:1184-7; PMID:16428425; http://dx.doi.org/10.1128/JB.188.3.1184-1187.2006.
37. Pope WH, Jacobs-Sera D, Russell DA, Peebles CL, Al-Atrache Z, Alcoser TA, et al. Expanding the diversity of mycobacteriophages: insights into genome architecture and evolution. PLoS One 2011; 6:e16329; PMID:21298013; http://dx.doi.org/10.1371/journal.pone.0016329.
38. Hendrix RW. Bacteriophages: evolution of the majority. Theor Popul Biol 2002; 61:471-80; PMID:12167366; http://dx.doi.org/10.1006/tpbi.2002.1590.
39. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proc Natl Acad Sci USA 1999; 96:2192-7; PMID:10051617; http://dx.doi.org/10.1073/pnas.96.5.2192.
40. Hendrix RW, Hatfull GF, Smith MC. Bacteriophages with tails: chasing their origins and evolution. Res Microbiol 2003; 154:253-7; PMID:12798229; http://dx.doi.org/10.1016/S0923-2508(03)00068-8.
41. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, et al. Origins of highly mosaic mycobacteriophage genomes. Cell 2003; 113:171-82; PMID:12705866; http://dx.doi.org/10.1016/S0092-8674(03)00233-2.
42. Bamford DH, Grimes JM, Stuart DI. What does structure tell us about virus evolution? Curr Opin Struct Biol 2005; 15:655-63; PMID:16271469; http://dx.doi.org/10.1016/j.sbi.2005.10.012.
43. Benson SD, Bamford JK, Bamford DH, Burnett RM. Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. Cell 1999; 98:825-33; PMID:10499799; http://dx.doi.org/10.1016/S0092-8674(00)81516-0.
44. Bamford DH. Do viruses form lineages across different domains of life? Res Microbiol 2003; 154:231-6; PMID:12798226; http://dx.doi.org/10.1016/S0923-2508(03)00065-2.
45. Abrescia NG, Bamford DH, Grimes JM, Stuart DI. Structure unifies the viral universe. Annu Rev Biochem 2012; 81:795-822; PMID:22482909; http://dx.doi.org/10.1146/annurev-biochem-060910-095130.

46. Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, et al. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet 2006; 2:e92; PMID:16789831; http://dx.doi.org/10.1371/journal.pgen.0020092.

47. Cresawn SG, Bogel M, Day N, Jacobs-Sera D, Hendrix RW, Hatfull GF. Phamerator: a bioinformatic tool for comparative bacteriophage genomics. BMC Bioinformatics 2011; 12:395; PMID:21991981; http://dx.doi.org/10.1186/1471-2105-12-395.

48. Pietilä MK, Laurinmäki P, Russell DA, Ko CC, Jacobs-Sera D, Butcher SJ, et al. Insights into head-tailed viruses infecting extremely halophilic archaea. J Virol 2013; 87:3248-60; PMID:23283946; http://dx.doi.org/10.1128/JVI.03397-12.

49. Burroughs AM, Iyer LM, Aravind L. Comparative genomics and evolutionary trajectories of viral ATP dependent DNA-packaging systems. Genome Dyn 2007; 3:48-65; PMID:18753784; http://dx.doi.org/10.1159/000107603.

50. Casjens S. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol 2003; 49:277-300; PMID:12886937; http://dx.doi.org/10.1046/j.1365-2958.2003.03580.x.

51. Katsura I, Hendrix RW. Length determination in bacteriophage lambda tails. Cell 1984; 39:691-8; PMID:6096021; http://dx.doi.org/10.1016/0092-8674(84)90476-8.

52. Pedersen M, Østergaard S, Bresciani J, Vogensen FK. Mutational analysis of two structural genes of the temperate lactococcal bacteriophage TP901-1 involved in tail length determination and baseplate assembly. Virology 2000; 276:315-28; PMID:11040123; http://dx.doi.org/10.1006/viro.2000.0497.

53. Siponen M, Sciara G, Villion M, Spinelli S, Lichière J, Cambillau C, et al. Crystal structure of ORF12 from *Lactococcus lactis* phage p2 identifies a tape measure protein chaperone. J Bacteriol 2009; 191:728-34; PMID:19047351; http://dx.doi.org/10.1128/JB.01363-08.

54. Xu J, Hendrix RW, Duda RL. Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. Mol Cell 2004; 16:11-21; PMID:15469818; http://dx.doi.org/10.1016/j.molcel.2004.09.006.

55. Haggård-Ljungquist E, Jacobsen E, Rishovd S, Six EW, Nilssen O, Sunshine MG, et al. Bacteriophage P2: genes involved in baseplate assembly. Virology 1995; 213:109-21; PMID:7483254; http://dx.doi.org/10.1006/viro.1995.1551.

56. Christie GE, Temple LM, Bartlett BA, Goodwin TS. Programmed translational frameshift in the bacteriophage P2 FETUD tail gene operon. J Bacteriol 2002; 184:6522-31; PMID:12426340; http://dx.doi.org/10.1128/JB.184.23.6522-6531.2002.

57. Bhattacharya B, Giri N, Mitra M, Gupta SK. Cloning, characterization and expression analysis of nucleotide metabolism-related genes of mycobacteriophage L5. FEMS Microbiol Lett 2008; 280:64-72; PMID:18248423; http://dx.doi.org/10.1111/j.1574-6968.2007.01047.x.

58. White MF. Homologous recombination in the archaea: the means justify the ends. Biochem Soc Trans 2011; 39:15-9; PMID:21265740; http://dx.doi.org/10.1042/BST0390015.

59. Yamtich J, Sweasy JB. DNA polymerase family X: function, structure, and cellular roles. Biochim Biophys Acta 2010; 1804:1136-50; PMID:19631767; http://dx.doi.org/10.1016/j.bbapap.2009.07.008.

60. Tanaka N, Shuman S. RtcB is the RNA ligase component of an *Escherichia coli* RNA repair operon. J Biol Chem 2011; 286:7727-31; PMID:21224389; http://dx.doi.org/10.1074/jbc.C111.219022.

61. Kaufmann G. Anticodon nucleases. Trends Biochem Sci 2000; 25:70-4; PMID:10664586; http://dx.doi.org/10.1016/S0968-0004(99)01525-X.

62. Phillips G, El Yacoubi B, Lyons B, Alvarez S, Iwata-Reuyl D, de Crécy-Lagard V. Biosynthesis of 7-deazaguanosine-modified tRNA nucleosides: a new role for GTP cyclohydrolase I. J Bacteriol 2008; 190:7876-84; PMID:18931107; http://dx.doi.org/10.1128/JB.00874-08.

63. Perreault J, Weinberg Z, Roth A, Popescu O, Chartrand P, Ferbeyre G, et al. Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. PLoS Comput Biol 2011; 7:e1002031; PMID:21573207; http://dx.doi.org/10.1371/journal.pcbi.1002031.

64. Goodrich-Blair H, Shub DA. Beyond homing: competition between intron endonucleases confers a selective advantage on flanking genetic markers. Cell 1996; 84:211-21; PMID:8565067; http://dx.doi.org/10.1016/S0092-8674(00)80976-9.

65. Palmer JR, Daniels CJ. In vivo definition of an archaeal promoter. J Bacteriol 1995; 177:1844-9; PMID:7896710.

66. Belcaid M, Bergeron A, Poisson G. The evolution of the tape measure protein: units, duplications and losses. BMC Bioinformatics 2011; 12(Suppl 9):S10; PMID:22151602; http://dx.doi.org/10.1186/1471-2105-12-S9-S10.

67. Bhaya D, Davison M, Barrangou R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 2011; 45:273-97; PMID:22060043; http://dx.doi.org/10.1146/annurev-genet-110410-132430.

68. Westra ER, Swarts DC, Staals RH, Jore MM, Brouns SJ, van der Oost J. The CRISPRs, they are a-changin': how prokaryotes generate adaptive immunity. Annu Rev Genet 2012; 46:311-39; PMID:23145983; http://dx.doi.org/10.1146/annurev-genet-110711-155447.

69. Terns MP, Terns RM. CRISPR-based adaptive immune systems. Curr Opin Microbiol 2011; 14:321-7; PMID:21531607; http://dx.doi.org/10.1016/j.mib.2011.03.005.

70. Zhang J, Kasciukovic T, White MF. The CRISPR associated protein Cas4 Is a 5' to 3' DNA exonuclease with an iron-sulfur cluster. PLoS One 2012; 7:e47232; PMID:23056615; http://dx.doi.org/10.1371/journal.pone.0047232.

71. Juhala RJ, Ford ME, Duda RL, Youlton A, Hatfull GF, Hendrix RW. Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages. J Mol Biol 2000; 299:27-51; PMID:10860721; http://dx.doi.org/10.1006/jmbi.2000.3729.

72. Cantalupo PG, Calgua B, Zhao G, Hundesa A, Wier AD, Katz JP, et al. Raw sewage harbors diverse viral populations. MBio 2011; 2; PMID:21972239; http://dx.doi.org/10.1128/mBio.00180-11.

73. Hatfull GF. The secret lives of mycobacteriophages. Adv Virus Res 2012; 82:179-288; PMID:22420855; http://dx.doi.org/10.1016/B978-0-12-394621-8.00015-7.

74. Arbiol C, Comeau AM, Kutateladze M, Adamia R, Krisch HM. Mobile regulatory cassettes mediate modular shuffling in T4-type phage genomes. Genome Biol Evol 2010; 2:140-52; PMID:20333230; http://dx.doi.org/10.1093/gbe/evq006.

75. Edgell DR, Gibb EA, Belfort M. Mobile DNA elements in T4 and related phages. Virol J 2010; 7:290; PMID:21029434; http://dx.doi.org/10.1186/1743-422X-7-290.

76. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 2007; 8:172; PMID:17521438; http://dx.doi.org/10.1186/1471-2105-8-172.

77. Lynch EA, Langille MG, Darling A, Wilbanks EG, Haltiner C, Shao KS, et al. Sequencing of seven haloarchaeal genomes reveals patterns of genomic flux. PLoS One 2012; 7:e41389; PMID:22848480; http://dx.doi.org/10.1371/journal.pone.0041389.

78. Godde JS, Bickerton A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol 2006; 62:718-29; PMID:16612537; http://dx.doi.org/10.1007/s00239-005-0223-z.

79. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 2002; 43:1565-75; PMID:11952905; http://dx.doi.org/10.1046/j.1365-2958.2002.02839.x.

80. Plagens A, Tjaden B, Hagemann A, Randau L, Hensel R. Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon *Thermoproteus tenax*. J Bacteriol 2012; 194:2491-500; PMID:22408157; http://dx.doi.org/10.1128/JB.00206-12.

81. Nuttall SD, Dyall-Smith ML. HF1 and HF2: novel bacteriophages of halophilic archaea. Virology 1993; 197:678-84; PMID:8249290; http://dx.doi.org/10.1006/viro.1993.1643.

82. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res 1998; 26:1107-15; PMID:9461475; http://dx.doi.org/10.1093/nar/26.4.1107.

83. Brenneis M, Hering O, Lange C, Soppa J. Experimental characterization of Cis-acting elements important for translation and transcription in halophilic archaea. PLoS Genet 2007; 3:e229; PMID:18159946; http://dx.doi.org/10.1371/journal.pgen.0030229.

84. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005; 33(Web Server issue):W244-8; PMID:15980461; http://dx.doi.org/10.1093/nar/gki408.

85. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 2009; 4:363-71; PMID:19247286; http://dx.doi.org/10.1038/nprot.2009.2.

86. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10; PMID:2231712.

87. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999; 27:573-80; PMID:9862982; http://dx.doi.org/10.1093/nar/27.2.573.

88. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000; 302:205-17; PMID:10964570; http://dx.doi.org/10.1006/jmbi.2000.4042.

89. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res 2004; 14:1188-90; PMID:15173120; http://dx.doi.org/10.1101/gr.849004.

90. Grissa I, Vergnaud G, Pourcel C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 2007; 35(Web Server issue):W52-7; PMID:17537822; http://dx.doi.org/10.1093/nar/gkm360.

91. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, et al. Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res 2011; 39(Database issue):D141-5; PMID:21062808; http://dx.doi.org/10.1093/nar/gkq1129.

92. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics 1998; 14:68-73; PMID:9520503; http://dx.doi.org/10.1093/bioinformatics/14.1.68.