Assembly of Viral Metagenomes from Yellowstone Hot Springs

Thomas Schoenfeld,[1*] Melodee Patterson,[1] Paul M. Richardson,[2] K. Eric Wommack,[3] Mark Young,[4] and David Mead[1]

Lucigen Corporation, Middleton, Wisconsin 53562,[1] *US Department of Energy Joint Genome Institute, Walnut Creek, California, 94598,*[2] *Department of Plant and Soil Sciences, University of Delaware, Newark, Delaware* 19711,[3] *Plant Sciences & Plant Pathology, Montana State University, Bozeman, Montana 59717*[4]

Thermophilic viruses were reported decades ago; however, knowledge of their diversity, biology and ecological impact is limited. Previous research on thermophilic viruses focused on cultivated strains. This study examined metagenomic profiles of viruses directly isolated from two mildly alkaline hot springs, Bear Paw (74°C) and Octopus (93°C). Using a new method for constructing libraries from picograms of DNA, nearly 30 Mb of viral DNA sequence was determined. In contrast to previous studies, sequences were assembled at 50% and 95% identity, creating composite contigs up to 35 kb, and facilitating analysis of the inherent heterogeneity in the populations. Lowering assembly identity reduced the estimated number of viral types from 1440 and 1310 to 548 and 283, respectively. Surprisingly, the diversity of viral species in these springs approaches that in moderate temperature environments. While most known thermophilic viruses have a chronic, non-lytic infection lifestyle, analysis of coding sequences suggests lytic viruses are more common in geothermal environments than thought. The 50% assembly included one contig of high similarity and perfect synteny to nine genes from Pyrobaculum spherical virus (PSV). In fact, nearly all the genes of the 28-kb genome of PSV have apparent homologs in the metagenomes. Similarities to thermoacidophilic viruses isolated on other continents were limited to specific open reading frames but were equally strong. Nearly 25% of the reads shared significant similarity between the hot springs, suggesting a common subterranean source. To our knowledge, this is the first application of metagenomics to viruses of geothermal origin.

Subterrestrial aquifers are vast ecosystems characterized by the absence of solar radiation, the presence of chemical reducing potential and, under certain conditions, elevated temperatures (31). Estimates of the volume of the global thermal aquifer range as high as 10^{19} liters (34), with microbial and viral abundances approaching those of the oceans (16). This study examined planktonic viruses directly isolated from two mildly alkaline siliceous hot springs in Yellowstone National Park (YNP). With temperatures of 74º and 93°C, life in these springs is comprised exclusively of *Bacteria*, *Archaea* and their respective viruses, all uniquely adapted to the temperature and chemistry extremes of the environment (65, 53). These springs are direct outflows of the thermal aquifer and not secondarily heated surface water (31). In this respect they are distinct from the acidic springs, mudpots, and other thermal features that have provided many of the published thermophilic virus samples. Conceivably, viruses may proliferate not only at the surface but deeper in the vent as well, where increased pressures and dramatically elevated temperatures have been measured. Water temperatures of 180-270°C are found at depths of 100-550 meters throughout the caldera of YNP (31). If viruses proliferate in the subsurface aquifer, hot springs separated by kilometer distances that share common water sources may also share viral populations.

Little is known about the roles of viruses in the ecology of hydrothermal environments, although they appear to play a role in host mortality and carbon cycling (16) and are probably the only predators. In better studied marine environments, an estimated 1030 viruses in the world's oceans (77) may comprise several hundred thousand different species (4). These viruses are responsible for a significant proportion of microbial mortality and thus have a profound influence on carbon and other nutrient cycles (77). Marine viruses are also thought to be important vehicles for lateral gene transfer via lysogeny and transduction and probably promote diversity by preferentially lysing the most abundant species (83). Analysis of viral metagenomes (4, 9, 21) and cultured viral genomes (43, 54) has consistently shown that about 30% of these sequences have detectable similarity to sequences in GenBank, and about half of these are most similar to other known viruses. In spite of extensive sequencing from oceanic phage and viral metagenomic samples, only small RNA genomes of 5-10 kb have been assembled (23) from virioplankton metagenomic sequence data.

Enrichment cultivation has been the primary tool for investigations of thermophilic viruses (those growing at >70°C). Since the first reports of thermophilic viruses (69, 47), hundreds of bacteriophages (88), dozens of crenarchaeal viruses (reviewed in 61, 74), and one euryarchaeal virus (32) have been isolated from thermal springs and vents around the world. Cultivated *Thermus* bacteriophages belong to four morphological families*: Myoviridae, Siphoviridae, Tectiviridae*, and *Inoviridae* (88). Their morphologies and the available genomic sequences (50, 38) suggest similarity to mesophilic bacteriophages. Most known thermophilic bacteriophages appear to be lytic, although this could be biased by the method of their discovery (88). Cultivated thermophilic crenarchaeal viruses infect the genera *Sulfolobus*, *Acidianus*, *Pyrobaculum*, and *Thermoproteus.* Morphologies and genome content suggest crenarchaeal viruses are unrelated to viruses of *Euryarchaeota, Bacteria* or *Eukarya* (57). All of the cultivated crenarchaeal viruses proliferate as chronic, nonlytic infections.

While enrichment cultures have been invaluable in the study of thermophilic viruses, important contextual information such as relative abundance, diversity and distribution is lost. Furthermore, these analyses exclude the majority of viruses that are not readily cultivated (73). No viral cultivation study fully replicates the temperature and pressure extremes and the chemistries that characterize the subsurface vents. Unlike cellular life, no universal genetic marker (e.g. rDNA) exists for viruses. Direct metagenomic analysis of viruses from environmental samples circumvents these limitations and provides insight into biology, evolution, and adaptations to the environment and composition of viral assemblages through studies of gene homology. No metagenomic analysis of water-borne viral populations in geothermal environments has been reported. In fact, planktonic life in thermal environments is under lored in general as microbial diversity studies of hot spring environments have focused almost exclusively on sediments (7, 10, 39), adherent filaments (64) or mats (82). The goal of this study was to profile the diversity, composition and adaptations of viral assemblages in two hot springs of YNP based on metagenomic analysis of viruses inhabiting these environments.

MATERIALS AND METHODS

Site description and sampling. Viral particles were isolated from Bear Paw (an unofficial name for LRNN374) (N 44.5560955 W110.8347866) and Octopus (N44.5340836 W110.7978895) hot springs (Table 1) (75). The temperatures of the hot springs are based on direct measurement on the day of

TABLE 1: Sample site and abundance of viral and bacterial counts

Hot Spring	Temp	pH	Cells/ml	Viruse/ml	Virus:Microbe Ratio	Virus/ml in Concentrate	Virus/ml Theoretical[a]	Efficiency
Bear Paw	74	7.34	$4.3x10^6$	$1.44x10^6$	0.33	1.48E+08	7.21E+09	2.1%
Octopus	93	8.14	$9.0x10^5$	$3.07x10^5$	0.34	2.18E+08	1.534E+09	14.2%

[a] Based on a concentration factor of 5000X (500L to 100mL).

the sampling. The pH values were determined by the USGS (48). Additional geochemical data for Octopus hot spring is available in Supplementary Table 1. Thermal water (400 – 600 l) was filtered using a 100 kD MWCO tangential flow filter (GE Healthcare). Viral particles were concentrated to 2 liters, filtered through a 0.2- µm filter and further concentrated to 100 ml using a 100-kD filter. Viral concentrates were imaged by transmission electron microscopy (TEM) (Leo 912AB operating at 80KV). Direct viral enumeration was performed by epifluorescence microscopy (51). Following the recommendations of Wen, et al. (84) samples were unfixed and were stained with SYBR Gold. The samples were stored at 4°C for no more than 24 hours before counting. Unfortunately, immediate snap freezing of samples in liquid nitrogen was not possible, so viral abundances may be somewhat underestimated.

Viral DNA processing and extraction. Viral concentrates were centrifuged at 12K rpm for 20 min., syringe-filtered using a 0.2-µm Acrodisc filter (Gelman) and further concentrated to 400 µl by filtration using a 30 kD MWCO Centricon spin filter (Millipore). Those judged by epifluorescence microscopy to be substantially free of microbial cells were used for library construction. Viral concentrates were transferred to SM buffer (0.1 M NaCl, 8 mM MgSO4, 50 mM Tris-HCl pH 7.5) using a 30 kD MWCO spin filter. Benzonase endonuclease (Sigma, 10 U) was added, and the reactions were incubated for 30 min. at 23ºC. EDTA (20 mM), SDS (0.5%) and Proteinase K (100 U) were added, and the reactions were incubated for 3 hours at 56ºC. NaCl (0.7M) and CTAB (1%) were added, and DNA was extracted with phenol/chloroform and ethanol precipitated.

Library construction and sequencing. Viral DNA was physically sheared to 3-6 kb using a HydroShear device (Genomic Solutions, MI). The ends were made blunt using the DNATerminator end repair kit (Lucigen, WI), and the fragments were ligated to a double-stranded asymmetrical linker comprised of one blunt phosphorylated end (5'-GATGCGGCCGCTTGTA TCTGATACTGCT-3', Linker 1) and one non-phosphorylated staggered end (5'-GGAGCAGTATCAGATACAAGCGGCCGCATC-3', Linker 2) to fix the primer in a defined orientation relative to the genomic DNA. Gel fractionation was used to remove unligated linkers and to isolate 3-6 kb fragments. These fragments were PCR amplified using Vent DNA polymerase (New England Biolabs, MA) and a primer targeted to Linker 1 (5'- AGCAGTATCAGA TACAAGCGGCCGCATC-3'). Amplification products were gel purified again, inserted into the cloning site of the transcription free pSMART vector (Lucigen) and used to transform E. cloni 10G cells (Lucigen). Libraries were sequenced by the Department of Energy's Joint Genome Institute (Walnut Creek, CA). The sequences were deposited in the GenBank trace archive and are retrievable using CENTER_NAME = "JGI" and SEQ_LIB_ID = "AOIX" for Bear Paw sequences and SEQ_LIB_ID = "APNO" and SEQ_LIB_ID = "ATYB" for octopus sequences.

Bioinformatics. Viral metagenome sequencing reads were compared to the nonredundant (nr) protein data base (GenBank) using BLASTx (2, 3). The fifty most significant BLASTx scores ($E < 10^{-3}$) were recorded. The occurrences of key words in the output of the BLASTx were counted using PERL scripts written for this project, and the sequences were categorized by function. Sequences were assembled using the SeqMan® program (DNAStar, WI) at a minimum of 50% or 95% identity over a minimum of 20 nt. Metagenome sequence libraries were compared to each other and to all the sequences in GenBank using tBLASTx (NBCI) with a cutoff of $e<10^{-3}$. Where indicated, the apparent open reading frames were identified and translated using the Gene Mark program (45). These translations were compared to the nr protein database using the BLASTp program. The rank abundances were calculated using the PHAge Community from Contig Spectrum (PHACCS) web utility located at http://phage.sdsu.edu/research/tools/phaccs/ (5).

RESULTS

Sampling sites, viral abundance, and morphologies. The two hot springs that provided samples are listed in Table 1. Bear Paw hot spring is in the river group of the lower geyser basin of YNP, while Octopus is about 5 km away in the White Creek area. The pH values of these hot springs are 7.34 and 8.14; the temperatures and apparent microflora differ widely. Bear Paw (74ºC) is characterized by orange sedentary microbial growth in the pool. Octopus water emerges at 93°C, the boiling point at the local elevation of 2300 meters, with none of the orange growth.

Virus enriched fractions were isolated from 400 to 600 L of hot spring water and concentrated by tangential flow filtration. Contaminating cellular DNA was greatly reduced by further filtration, centrifugation and nuclease treatment. Viral and microbial direct counts of concentrated and unconcentrated samples were used to determine initial abundance, filtration efficiency and virus to microbe ratio (VMR). Viral abundance ranged from about 105 to 106 viruses ml-1 147 (Table 1), which is at the lower end of the range of 104 to 109 148 reported for thermal springs in California (16) and moderate temperature aquatic environments (86). The VMRs in YNP hot springs were nearly identical at 0.3, much lower than in moderate temperature environments (typically 3 to 10). These low VMRs may be related to the observation that none of the cultured thermophilic crenarchaeal viruses proliferate via lytic infections, a lifestyle that would result in large burst sizes at the same time as the microbial population is reduced. Two and 14% of theoretical yields of viruses in the two hot springs were isolated in the concentrates. It is not known if this loss was systematic and, therefore, biased the metagenomic analysis. Tailed, rod shaped, and filamentous morphologies were observed in the concentrates (Figure 1). Morphologies of viral particles in the concentrates represent most morphological families of known thermophilic viruses. Spherical viruses, as well as the other morphotypes, have been seen in direct concentrates from other similar hot springs (data not shown). Tailed morphologies are commonly associated with bacteriophages and euryarchaeotal viruses (88, 32); rod-shaped and filamentous morphologies are more commonly associated with crenarchaeal viruses (58).

Library construction and sequencing Advances in sequencing capacity make analysis of large numbers of clones feasible; however, challenges in sampling and library construction have prevented the widespread use of metagenomic shotgun sequencing for studying viral populations. At around 50 attograms of DNA per virus, abundances of 10^5 - 10^6 virus per ml correspond to 5 - 50 ng of viral DNA per liter. In practice, processing of hundreds of liters of spring water yielded at

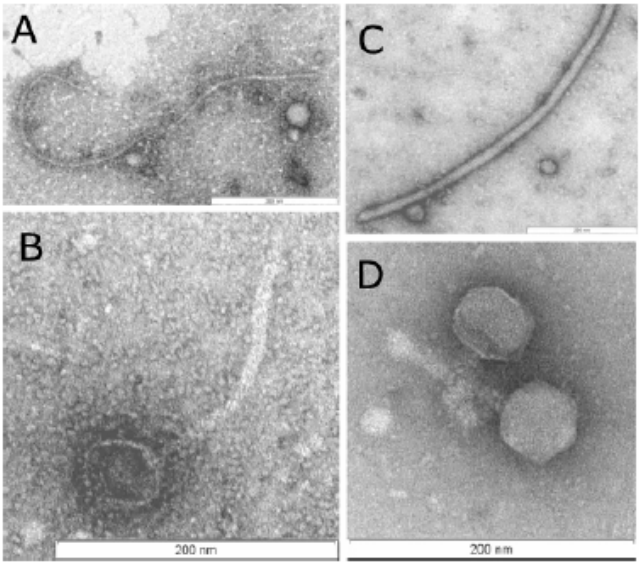

FIG. 1.. TEM Images of virus-like particles directly isolated from YNP hot springs. Images from Bear Paw (Panels A and B) and Octopus (Panels C and D) hot springs are shown. The bar in each figure is 200 nm.

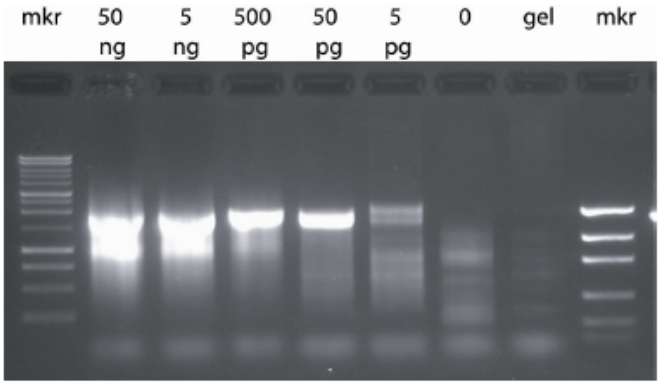

FIG. 2. Sensitivity of linker-facilitated anonymous DNA amplification. Decreasing amounts of lambda DNA (50 ng, 5 ng, 500 pg, 50 pg and 5 pg, as indicated) were sheared, end repaired, linkerligated, and size selected on an agarose gel. The resulting material was amplified using Vent DNA polymerase and a single oligonucleotide complimentary to the linker sequences. One tenth of the reaction was resolved by agarose gel electrophoresis as shown. As negative controls, no input DNA (lane 0) or only a DNA-free gel slice from otherwise identical reactions were similarly processed. DNA molecular weight markers (mkr) are indicated.

most 100 ng of DNA, much lower than is normally required for library construction. This low yield of virus precluded cesium chloride purification of the viral particles, as is commonly used for viral metagenomic library construction. Viral DNA also contains cytotoxic genes and modified nucleotides that induce host restriction systems. A new linker-dependent, anonymous method of DNA amplification was used to access this diversity, allowing construction of 3-8 kb insert libraries with none of the potential modified nucleotides. Viral DNA was physically sheared and short (20bp) linkers were ligated to the DNA fragments to serve as priming sites for PCR. Amplified fragments were cloned into a transcription-free pSMART vector to minimize cloning bias due to cytotoxic sequences (33). The use of flanking synthetic linkers provides identical primer annealing sites for each viral template in the mixture, which significantly limits amplification bias.

Before use on the thermophilic viral DNA, the amplification and library construction methods were validated using lambda DNA as a model template. After linker ligation and processing, as little as 5 pg of DNA was amplified and cloned (Figure 2). Comparison of 50 clones to lambda whole genome sequence showed twenty-two discrepancies among 39,000 bases sequenced, an error rate of 0.05%. Coverage was 64.5% with no apparent stacking of reads.

A total of 28,883 sequence reads were determined from Bear Paw (7685 reads) and Octopus (21,198 reads) hot springs. Paired-end reads averaged 981 nucleotides each, or nearly 30 Mb total. Assuming an average genome size of 50 kb, which is supported by agarose gel electrophoresis of the viral genomic DNA (data not shown), this sequencing depth represents about 600 viral genomic equivalents. The quality of the libraries is highly dependent on the amount of DNA used in their construction. The sequence reads of the Octopus library contained very few anomalies that would suggest amplification bias or cloning artifacts. Some of the reads from the Bear Paw library were less random than the Octopus library, as determined by several cases of sequence stacking. This library construction method has been used in the analysis of several cultivated and uncultivated viral genomes (9, 13, 14, 15, 44, 52, 70) but has never been fully described. It is being described in detail here for the first time.

Contaminating cellular DNA in viral DNA preparations was greatly reduced by filtration and nuclease treatment. Only viral preparations substantially free of microbial cells as judged by epifluorescence microscopy were used for library construction. Detection of rDNA sequences (5S, 16S, and 23S) in the libraries was used to identify contaminating cellular DNA. These sequences are absent in known viral genomes but highly conserved in microbial cells. A typical bacterial genome contains 15 rRNA genes (22). Most hyperthermophilic archaeal and bacterial genomes contain 3 or 6 rRNA genes, although the genomes of certain moderately thermophilic Geobacillus that grow at the temperature of Bear Paw contain up to 30 rRNA genes (26). BLASTn analysis identified only four rDNA sequences in the 10.4 microbial genome equivalents sequenced from the Octopus library (two 23S and two 16S), and eight in the 3.8 microbial genome equivalents from the Bear Paw library suggesting viral enrichment was quite high, particularly for the Octopus library. This inference is supported by a high similarity to sequences of cultivated viruses (shown below) and a large number of BLASTx similarities to genes associated with viral functions. In particular, the hundreds of presumptive genes for viral functions, such as replication, transcription, translation, lysogeny, recombination, lysis and structural proteins (Table 2, Supplementary Table 2) is consistent only with a predominately viral origin of the sequences.

Identification of likely gene products. BLASTx analysis of the individual reads was used to infer function of the sequences in the libraries. While most reads revealed no significant similarity to known proteins, both of the libraries contain numerous similarities to genes encoding proteins unique to viruses and phages (Table 2 and Supplementary table 2). Certain similarities were particularly interesting. Hundreds of contigs showed sequence similarity to the superfamily II helicases of a wide range of cells and viruses. For example, the 2-kb Octopus contig 158 had significant similarity to helicases of bacterial, archaeal and eukaryotic cells, as well as to phage and archaeal viruses. Species with significant expect values included *Staphylococcus phage Twort* (2E-16), *Myxococcus xanthus* (1E-15), *Sulfolobus islandicus filamentous virus* (8E-15), *Pyrococcus abyssi* (4E-08), *Eremothecium gossypii* (a fungus, 9E-05), *Tribolium castaneum* (an insect, 4E-04) and *Homo sapiens* (6E-03). Although lysin genes were highly abundant and are typically proximal to holin genes, no homologs for holins were seen, probably due to the high molecular diversity observed in known holin genes (81).

Sequence assembly and estimation of viral diversity. The degree to which metagenomic reads assemble has been used to assess the diversity of the viral populations. Most previous studies have used 98% identity over 20 nucleotides as the assembly stringency (e.g. 4, 13, 14, 15). In this study, sequence reads were assembled at ≥95% and

TABLE 2: Functional categories based on key word analysis of the BLASTx similarities

	Number of Reads Matching a Keyword		Percent with a Keyword	
COGs functional category	Bear Paw	Octopus	Bear Paw	Octopus
No BLASTx similarity	2545	8469	-	-
F. Nucleotide transport and metabolism	1445	2130	35.09%	37.81%
J. Translation, ribosomal structure and biogenesis	221	336	5.37%	5.96%
K. Transcription	278	325	6.75%	5.77%
L. Replication, recombination, and repair	688	989	16.71%	17.55%
O. Posttranslational modification, protein turnover, chaperones	181	213	4.40%	3.78%
None – virus-specific	350	596	8.50%	10.58%
No match to a keyword	955	1045	23.19%	18.55%

TABLE 3. Sequence assembly data and estimation of viral diversity

	Bear Paw	Octopus		
Sequence reads	7,685	21,198	28.883	
	Bear Paw 95%	Octopus 95%	Bear Paw 50%	Octopus 50%
Contigs assembled	6,191	13,543	4,850	4,788
Avg. reads per contig	1.239	3.129	1.587	4.427
Largest contig (nt)	3,503	4,554	8,007	35,089
Power law richness	1,440	1,310	548	283
Evenness score	0.946	0.954	0.933	0.936
Most abundant virus	2.14%	1.88%	3.93%	4.88%
Shannon-Weiner score	6.88	6.85	5.88	5.29

≥50% identity over a minimum match length of 20 nucleotides (Table 3). Diversity was estimated at both assembly stringencies using the power law rank-abundance model built into the Phage Communities from Contig Spectrum tool (PHACCS) (5) using an average genome size of 50 kb. As expected, diversity estimates were highly dependent on the assembly parameters (Table 3). The power law model predicted 1400 and 1310 viral types in Bear Paw and Octopus hot springs, respectively, at the 95% identity assembly level but decreased to 548 and 283 types, respectively, at 50% stringency.

Although assemblies at 95% identity or higher are typically used for metagenomic studies, assemblies at 50% are quite useful for studying diversity among related viruses. At 95% identity, the largest contigs were 3.5 and 4.6 kb for Bear Paw and Octopus, respectively (Table 3). At 50% identity, Octopus reads assembled into 17 contigs of greater than 10 kb, including contigs of 35 kb and 19 kb, comprised of >1000 reads each (Supplementary Table 3). In each case, reads were evenly distributed across the contigs. The four strongest BLASTx hits to the 35-kb contig belonged to thermophilic crenarchaeal viruses *Acidianus Rod-shaped virus* (ARV), *Sulfolobus islandicus rod-shaped viruses* 1 (SIRV1) and 2 (SIRV2), and *Sulfolobus islandicus filamentous viruses* (SIFV) (Table 4). The only significant similarity for the 19 kb contig was to the thermophilic crenarchaeal virus, *Pyrobaculum spherical virus* (PSV). The seventeen contigs >10 kb comprise a total of 7 Mbp of sequence, or 140 viral equivalents. In the Bear Paw library, with roughly one third as many reads, the largest contig that assembled at 50% identity was 8 kb. Five hundred thirty four (7%) of the reads assembled into 19 contigs of >4 kb. These include 0.5 Mbp of reads, or 10 viral equivalents.

Certain contigs provide compelling evidence that the 50% assemblies associate genuine orthologous sequences. An example is Bear Paw contig 327 (Figure 3). Eleven open reading frames (ORFs) were identified by the GeneMark algorithm (45). BLASTp analysis of each shows strongest similarity to the putative coding sequences of PSV (36). Nucleotide identities were as high as 88%, gene order is perfectly preserved relative to the cultured virus, and gene overlap is identical between the composite contig and the cultivated virus. Interestingly, two different ORFs of the PSV genome, gp 4 and 5 are apparently related to each other, since both had significant similarity to the same region of the consensus contig. In both the cultured viral

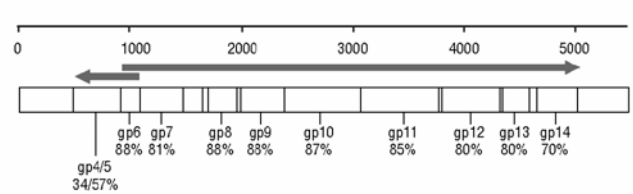

FIG. 3**.** Genes and gene order are highly conserved between a cultured crenarchaeal virus and a consensus contig from the Bear Paw library. Contig 372 (5492 bp, 71 reads) was assembled at ≥50% identity from the Bear Paw library. Open reading frames identified by GeneMark algorithm were compared by BLASTp to proteins in GenBank. Similarities to *Pyrobaculum spherical virus* proteins are shown with percent coding identity. The gene names are based on the annotation in GenBank and are named in order of their location on the viral chromosome. Direction of transcription is indicated by the arrows.

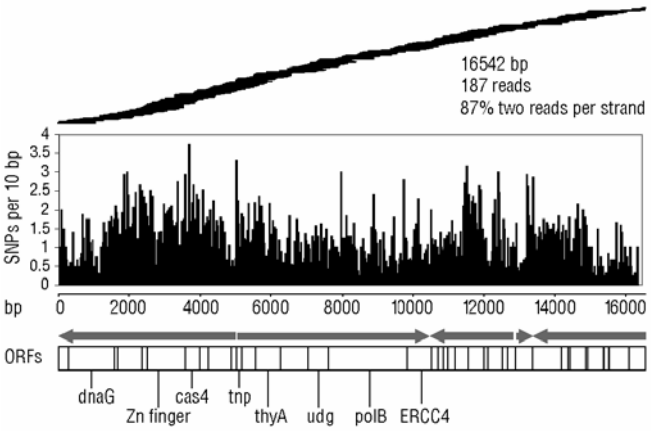

FIG. 4. Alignment of nucleotide polymorphisms with coding sequences in a 16.5 kb consensus contig from Octopus Hot Spring. Contig 722 was assembled at ≥50% identity from the Octopus library. Sequence coverage is shown on the top, with each line representing a separate read. Single nucleotide polymorphisms per 10 base pairs were normalized to the number of reads covering the respective nucleotide (middle) and are aligned with predicted open reading frames from the consensus sequence in the contig and the gene name of the strongest BLASTx similarity (bottom). Direction of transcription is shown by the arrows. Similarities to known genes were identified by BLASTp.

genome and the consensus contig, the gp7 PSV gene overlaps gp6 in the opposite orientation.

Contig 722 from the Octopus spring library provided a unique opportunity to associate population diversity of an assembled metagenome with the biochemistry of the gene products (Figure 4). This 16.5 kb contig, assembled at 50% identity, includes 187 reads (average coverage of 11 reads per nucleotide position). GeneMark predicted 26 ORFs of greater than 100 nucleotides, including an apparent replication operon. The genes with the strongest similarity to four of these ORFs encode primase, uracil DNA glycosylase, family B

TABLE 4. Numbers of 95% contigs with tBLASTx similarities to cultured viral sequences

			Number of TBLASTx similarities	
Virus	Reference	Accession	Bearpaw	Octopus
ARV, *Acidianus rod-shaped virus*	79	AJ875026	36	228
SIRV 1 & 2, *Sulfolobus islandicus rod-shaped virus 1 & 2*	11,55	AJ344259 AJ414696	30	217
PSV, *Pyro bacalum spherical virus*	36	AJ635161	44	152
SIFV, *S. islandicus filamentous virus*	6	AJ440571	7	46
STSV1, *S. tengchongensis spindle-shaped virus 1*	87	AJ783769	26	22
ATV, *Acidianus two-tailed virus*	62	AJ888457	8	17
YS40, *Thermus thermophilus YS40 phage*	50	DQ997624	15	41
TTSV1, *Thermoproteus tenex spherical virus 1*	1	AY722806	6	12
Twort, *Staphylococcus phage Twort*	43	AY954970	4	21

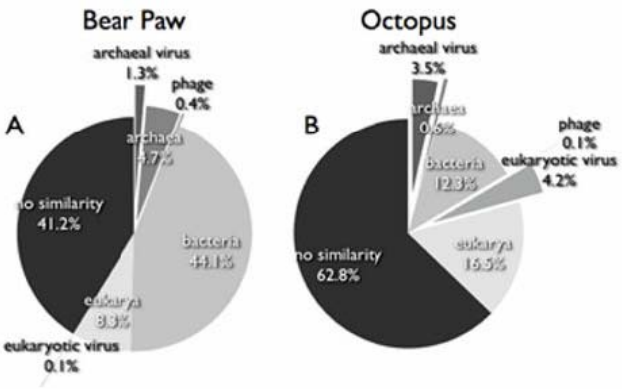

FIG. 5. Broad classification of viral metagenomic contigs based on tBLASTX similarities. Contigs assembled at 95% identity from Bear Paw and Octopus reads (Panel A and B, respectively) were compared to sequences in GenBank to infer phylogeny. Shown are frequencies of contigs with no significant sequence similarity in GenBank (E<0.001) and those with sequence similarity to *Bacteria*, *Archaea*, *Eukarya* and their respective viruses.

DNA polymerase, and nucleotide excision repair nuclease (*dnaG*, *udg*, *polB* and ERCC4 genes, respectively). Homologs of these ORFs belong to crenarchaeal DNA replication/repair complexes (8, 25, 68). The predicted *polB* gene showed 28% identity to *Pyrobaculum islandicus* polB2 (41), and has an archaeal codon profile (data not shown). Sequences from three of the discreet clones that comprise the *polB* gene in this contig have been expressed in *E. coli* to produce a functional thermostable DNA polymerase (data not shown). This contig also contains apparent homologs to a zinc finger-like protein and a transposon-like integrase/resolvase (*tnp*), functions commonly associated with viruses and phages. Another ORF with highest similarity to the CRISPR-associated sequence *cas*4 (35) is unlikely to be part of a functional CRISPR system. Unlike authentic Cas sequences, this one is virus-derived and is not proximal to a CRISPR sequence or other typically associated sequences. More likely this gene is a separate member of the Cas4 COG, presumably a RecB-like exonuclease (35).

To correlate the level of sequence divergence with predicted gene function, SNP frequency was aligned to the 50% assembly consensus sequence of the contig. Overall distribution of SNPs in the contig was 0.705 per 10 bp. Replication-associated genes showed noticeably lower molecular diversity than the other ORFs. SNP distribution in the *dnaG*, *udg*, *polB* and ERCC homologs was 0.565, 0.617, 0.569 and 0.548 per 10 bp, respectively, while the distribution in the Zn finger, *cas4* and *thyA* homologs was 0.979, 1.31, 0.728, respectively.

Similarities to known viral and microbial genomes imply phylogeny. tBLASTx analysis was used to infer phylogenetic origin of the 95%-assembled contig sequences. A majority of the contigs (41% from Bear Paw and 63% from Octopus) had no tBLASTx similarity (E<0.001) to any sequence in GenBank (Figure 5). Although it is typical for viral metagenomic libraries analyzed in this way to have a high proportion of sequences without identifiable homologs, these libraries contained the highest frequency of novel sequence reported to date. This trend likely reflects the lack of genetic sequence data from microorganisms in high temperature environments. The libraries were noticeably different from one another with regard to the frequency of reads within each of the seven tBLASTX homology groups, The Bear Paw library had a 4 to 5-fold higher frequency of bacterial and archaeal sequence similarities (44 and 5%, respectively) than the Octopus spring library (12 and 1%, respectively). It is tempting to speculate that this reflects a higher relative abundance of bacteriophage at the lower temperature; however, this may also be related to a potentially higher level of microbial DNA contamination in the Bear Paw library indicated by rRNA sequences (above).

Interestingly, the libraries contained a sizable number of sequences with homology to eukaryotic genes, 16.5% for Octopus Spring and 8.3% for Bear Paw, which may reflect the commonly

TABLE 5. Numbers of 95% contigs with tBLASTx similarities (E<0.001) to the respective cellular genome

Archaea	Bear Paw	Octopus
Pyrobacalum	124	684
Aeropyrum	62	626
Sulfolobus	38	326
Acidianus	25	185
Bacteria		
Aquifex	474	1138

observed overlap in gene sequence homology between *Archaea* and *Eukarya* in general (18). Almost all known crenarchaeal viruses infect three archaeal genera, *Pyrobaculum, Sulfolobus* and *Acidianus.* Interestingly, these genera were three of the four most common archaeal sources of the sequence similarities to the two libraries, the other being *Aeropyrum* (Table 5). Genetic similarities to *Sulfolobus* and *Acidianus* are surprising because these two genera are found exclusively in highly acidic environments. Nearly half the bacterial similarities were to *Aquifex*. To our knowledge, no attempts have been made to cultivate phage on any strain in the *Aquificales* order.

Overall, only 3.4% of the 95% contigs from the two libraries showed similarity to known viral sequences. Most of these similarities were to cultivated thermophilic crenarchaeal viruses (Table 4). Similarity to the only non-thermophilic virus, *phage Twort* (43), was limited to the helicase gene, which shares similarity with that of SIFV (see above). The two libraries shared similar frequencies of genetic homology to archaeal viruses and phage. Notable exceptions were *Acidianus rod-shaped virus* and *Sulfolobus islandicus rod-shaped* virus 1& 2 where the Octopus library demonstrated a higher frequency of homology than the Bear Paw library and the *S. tengchongensis spindle-shaped virus 1* homology, less common in Octopus than in Bear Paw (Table 4). Alignment of the metagenomes to whole genome sequences of six cultivated thermophilic viruses revealed striking conservation of certain sequences (Figure 6). Almost the entire genome of *Pyrobaculum spherical virus* has similarity to sequences in both metagenomic libraries, with median identities of 60% and 51% to the Bear Paw and Octopus 95%- contigs, respectively. Sequence similarities to the other crenarchaeal viruses and to bacteriophage YS40 were limited to a few specific ORFs, but the degree of similarity was relatively high in those regions. Interestingly, nearly all of the ORFs showing high levels of homology are among the few thermophilic crenarchaeal virus genes for which a function has been assigned or inferred (Figure 6 and references therein). These regions of high conservation are genes associated with virion components, DN310 A replication, transposition, recombination or nucleic acid metabolism.

Similarities between the two hot springs viral populations. Nucleotide sequences within the two libraries were compared to one another to assess similarity between the viral populations in the two very different thermal environments. Contigs assembled at 95% from the two libraries were compared to each other by tBLASTx and BLASTn (Table 6). The differences between the two analyses should be the result of non-coding nucleotides. Since gene densities are high in viral genomes and there is very little intergenic sequence, these differences are mainly due to silent codon variations, which should be largely free of selective pressure. Most remarkable is the high degree of similarity between the two libraries by either analysis. By tBLASTx, 5,843 of the Octopus contigs (43%) and 1,593 of the Bear Paw contigs (26%) shared amino acid coding similarity. By BLASTn, 2876 (21%) and 1339 (21%) of the respective contigs shared nucleotide similarity. The average percent identities were 74% and 87% and the expect values were 1.38E-05 and 3.00E-05, although the average length of sequence alignment (298 and 175 bp) was modest in both cases. This level of similarity did not allow extensive assembly of contigs from the two libraries, even at 50% identity, presumably due to the short lengths of alignment (not shown). Taken together, these data suggest a mosaic-like pattern of overlap of much of the coding content in the two hot springs, although entire viral genomes or even entire genes are not necessarily fully conserved. The fact that the degree of identity at the

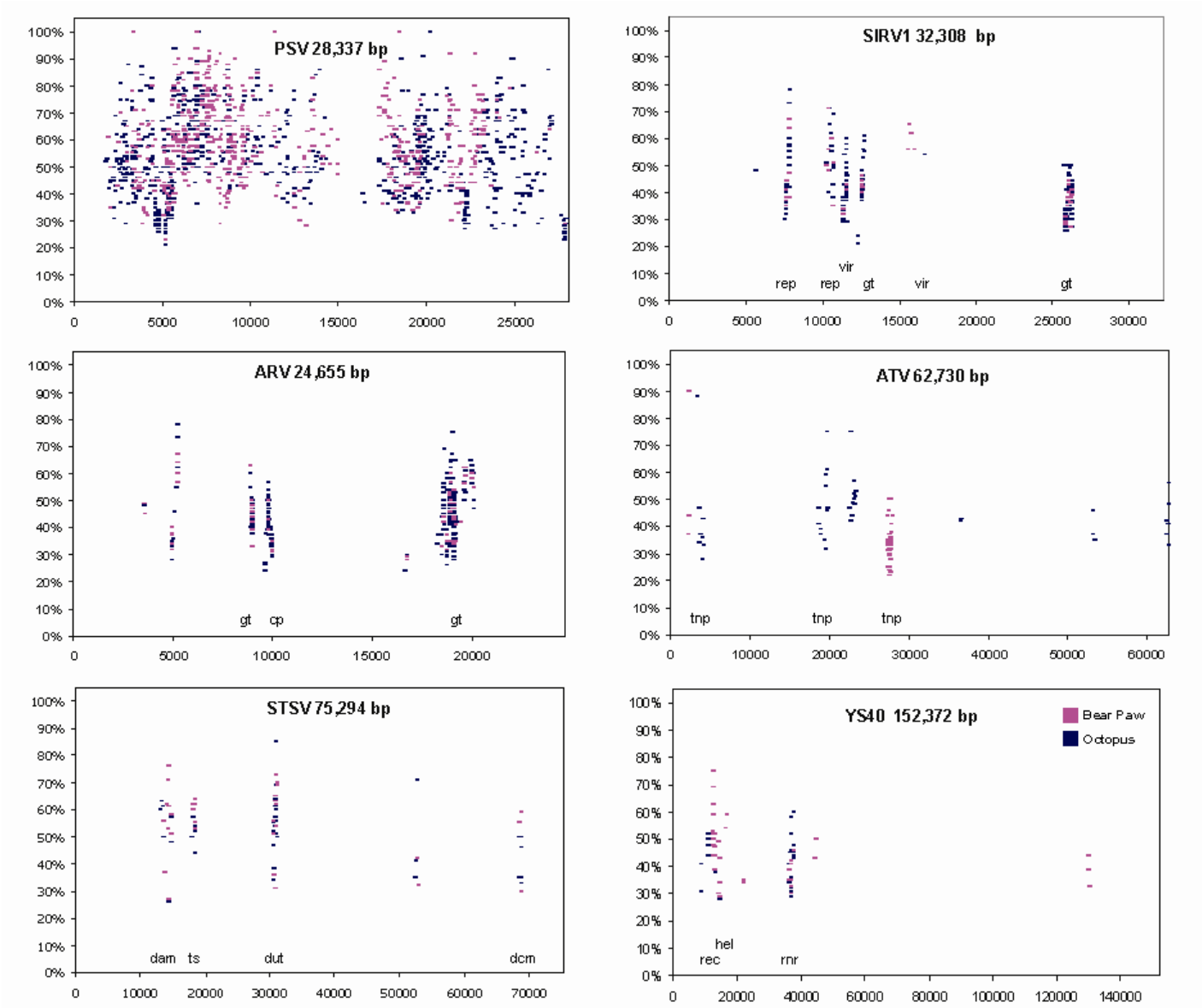

Fig. 6. Alignment of Octopus and Bear Paw viral metagenomic library contigs with six cultured virus genomes. Contigs assembled at >95% identity from the viral metagenomic libraries were compared by tBLASTx to the genomes of PSV, SIRV1, ARV, ATV, STSV and YS40. Each bar represents the alignment of a unique metagenomic sequence to the indicated location on the cultivated viral genome, shown on the horizontal axis. Percent coding sequence identities are shown in the vertical axis. The threshold for inclusion of a contig is an E-value <10-3. Red bars indicate Bear Paw alignments; blue bars indicate Octopus alignments. Also shown are the known or predicted functions of the conserved coding sequences (*rep*, replication related; vir, virion component; *gt*, glycosyltransferase; *tnp*, transposase; cp, coat protein; *dam*, adenine DNA methylase; *ts*, thymidylate synthase; *dut*, dUTPase; *dcm*, cytosine DNA methylase; *hel*, helicase; *rec*, recombinase; *rnr*, ribonucleotide reductase (1, 6, 11, 36, 42, 50, 55, 62, 79, 87 and D. Prangishvili, personal communication).

nucleotide level and at the translational level were relatively close suggests that this overlap is not due solely to selective pressure on the coding sequence, but must be explained by other mechanisms.

TABLE 6. Nucleotide and coding similarities between viral populations of Octopus and Bear Paw Hot Springs

	Bear Paw	Octopus
Frequency (number) of Octopus contigs with similarity to Bear Paw contigs	43% (5843)	21% (2876)
Frequency (number) of Bear Paw contigs with similarity to Octopus contigs	26% (1593)	21% (1339)
Average length of similarity (nt)	298	175
Average identity	74%	87%
Average expect value	1.38E-05	3.00E-05

DISCUSSION

Viral diversity in geothermal environments. Octopus hot spring is well documented to support prolific microbial life (17) and its geochemistry (48, Supplementary Table 1) is suitable for chemotrophic lifestyles. Analyses based on rDNA sequences (10, 64) show that microbial diversity is relatively limited compared to moderate temperature environments. These studies and studies of lipid and isotope composition (40) suggest the microbes in the filaments and the sediments, close in proximity and temperature to the sample site in this study, are primarily *Bacteria*, with *Aquificales* and *Thermatogales* most highly represented. The mats in the Octopus outflow that have been studied in detail (82) are in much cooler regions several meters away from the sample site and have very different microbial populations. To our knowledge, no detailed study of the chemical composition or life in Bear Paw has been published. In fact, no study of planktonic life or population-level analysis of viral assemblages has been reported for any thermal environment.

Since this study suggests that hundreds or thousands of different viral genotypes probably inhabit the two thermal springs and more than

half of the sequences collected have no apparent function or database homolog, much of the overall genetic diversity remains to be described. Metagenomic analysis is a useful tool for accessing this diversity and complements the cultivation-based research. A limitation of this approach is that it only allows analysis of dsDNA viruses. All cultivated thermophilic viruses have dsDNA genomes except certain *Thermus*-specific *Inoviruses*, which have ssDNA genomes (88). Notably, several viral nucleic acid preparations from this study had RNase-digestible material (data not shown), suggesting that RNA viruses inhabit these hot spring environments.

Assembly of metagenomic sequences has been the only available means of assessing molecular diversity in viral populations (15, 5). Since most of these previous studies have used assemblies at 95% or higher, the 95% assemblies are useful for comparing the hot springs viruses described here to other environments. Based on average genome sizes of 50 kb, the power law model of the PHACCS tool (5) estimated species richness at 1440 and 1310, respectively, for Octopus and Bear Paw hot springs with no viral species representing more than 2% of the total population (Table 3). These estimates are similar to the 1,650 viral genotypes estimated for Chesapeake Bay virioplankton using an average viral genome size of 125 kb (9); but, are only roughly one half to one quarter the diversity of marine environments and fecal samples which were modeled using an average viral genome size of 50 kb (4, 13, 14, 15). This model gave estimates of 3350 to 7180 viral genotypes in the marine planktonic samples, 7340 in marine sediment, and 2390 in human feces and indicated that no single viral species represents more than 2-3% of the total population in the open ocean, 0.014% in the marine sediment, or 4.8% in human feces.

There are several limitations in assessing actual numbers of viral species from metagenomic libraries. First, these models assume viral genomes evolve uniformly. However, different regions of viral genomes are clearly more conserved than others (Figure 4; and ref. 46). Genetic diversity outside the conserved regions is probably higher than these models indicate. A second complication is that classical species definitions are poorly suited to viruses in general (19). Cultured viruses can be grouped by host range, morphology, replication lineages and physicochemical and antigenic properties. Metagenomic sequences cannot be associated with these criteria. Assembly at >95% nucleotide identity fails to account for molecular diversity among related viral types, which is higher than that of cellular species. In fact such stringency would fail to associate viruses that, based on the above criteria, are known to be related (37, 43, 46) although they may share as little as 50% nucleotide identity over much of their genomes. Finally, the generation of new viral species by mosaicism, modular evolution, or lateral gene transfer (20, 80, 83) would not be detected using assembly of <1 kb sequence reads. On the other hand, given the dynamic nature of viral genomes, this approach is well-suited to a view of the diversity and evolution of viruses that considers genes or groups of genes rather than whole genomes.

Genome heterogeneity revealed by lower stringency assemblies. Assembly at 50% identity allows for intrinsic sequence heterogeneity among related viral types. These assemblies appear very reliable in associating orthologous sequences. Particularly in the Octopus library, the sequence reads are evenly distributed throughout the contigs with minimal stacking or other anomalies that would suggest amplification or cloning artifacts. The high numbers of reads on both strands, evenly distributed throughout the contigs, suggest these contigs represent independent clones of closely related genomes. Contig 372 from the 50% assembly of Bear Paw reads provides compelling evidence that lower stringency assembly associates related sequences. This 5.5 kb contig contains 11 open reading frames, seven of which share between 72 and 88% nucleotide identity to the cultivated PSV genome, as well as gene order and overlap of ORFs.

It is likely that some of the larger contigs represent nearly complete consensus genomes. Most of the cultivated crenarchaeal viral genomes are 35 kb or less (61); the *Sulfolobus* spindle-shaped viruses are as small as 15 kb (85). The two largest assemblies at 50% identity, 35 and 19 kb, had BLASTx similarities to only thermophilic crenarchaeal viruses. These two contigs alone represent 7.5% and 5.2% of the total reads in the Octopus library, slightly above the power law prediction that no single species is greater than 4.88% of the total viral population. Since the seventeen >10 kb contigs represent 29% of the Octopus reads, the sequences determined in this project appear highly representative of the most abundant viruses in Octopus hot spring.

Using the lower stringency assemblies, SNPs can be identified and mapped to the coding sequences. Some genes of the Octopus contig 722 (Figure 4) assembled at 50% identity have been expressed to produce active multisubunit DNA polymerases (data not shown), confirming the homology to the *udg*, *polB* and ERCC4 operon suggested by BLAST alignment. This replication operon is more conserved than the surrounding regions. As additional biochemical and structural data become available, molecular diversity may be correlated with variations in function and structure.

As is common for viral populations (15, 21) and cultured phages (54), most sequences had no similarity to known genes, which probably reflects the overall diversity of viral genes and the relative paucity of annotated viral sequences. Nearly all noted similarities were to bacterial and archaeal genes that that are known to have viral counterparts (e.g. helicases, ribonucleotide reductases, thymidylate synthases and DNA and RNA polymerases).

Viral lifestyles and the role of viruses in lateral gene transfer in thermal environments. The identification of certain genes allows insights into viral lifestyles. For example, 532 lysin-like genes among 600 viral equivalents suggests lytic viruses are quite common in the hot springs, in contrast to the cultured thermophilic crenarchaeal viruses, all of which are nonlytic. The 86 apparent integrase genes imply that lysogeny is also common in thermal aquifers, consistent with previous studies that show integrase homologs in six crenarchaeal viral genomes (ATV, STSV1 and four SSV isolates) (62, 85, 87), and induction of prophage by mitomycin C in 1-9% of hot spring microbial cells (16).

Viruses have been implicated in lateral gene transfer and nonorthologous gene replacement in cellular genomes (24, 80) and may have played critical roles in the evolution of DNA and DNA replication mechanisms, the separation of the three domains of life and the origin of the eukaryotic nucleus (reviewed in 29). Gene similarities seen in the metagenomic libraries (Supplementary table 2) support the role of viruses in cellular evolution. Similarities to reverse transcriptases were almost exclusively to the intron associated maturase/reverse transcriptases and retrotransposon reverse transcriptases. These genes and the numerous recombinase, integrase, and transposase genes suggest that appropriate machinery for lateral gene transfer exists in hot spring viral genomes (20). Other gene homologs provide evidence of ongoing gene transfer within these populations. Helicase genes shared among viruses and cells from all domains have been considered examples of nonorthologous replacement of cellular genes by viral genes (28). Helicase genes similarities were abundant in the viral libraries (117 and 217 in Bear Paw and Octopus, respectively) and many of these had high similarity to cellular and viral genes from all three domains of life. Also common in the metagenomic libraries are presumptive ribonucleotide reductases (14 and 50, respectively) and thymidylate synthase (7 and 51, respectively) genes. The conservation these genes in viral and cellular genomes of all domains and the biochemical activities of the gene products imply that viral genes played a key role in the transition from RNA-based to DNA-based genomes (30). DNA polymerase (*pol*) genes have also been proposed as likely examples of nonorthologous replacement by viral genes (27). Over two hundred apparent *pol* gene homologs were identified in the two metagenomic libraries, with all the polymerase families represented. Some of the viral *pol* gene products have unique biochemical properties that will be described elsewhere. In contrast, no *pol* gene has been identified by BLASTx analysis of the known crenarchaeal viral genomes, and *pol* genes have been reported in only two thermophilic bacteriophage genomes (38, 50). In addition to the *lys* genes, the high abundance of *pol* genes in the metagenomic libraries compared to cultured genomes suggests that our view of diversity may be biased by the difficulty in culturing certain types of viruses.

Origins and distribution of viruses in thermal springs. Metagenomic analysis allows insight into the global and regional

distribution of viruses. Alignment of metagenomic sequences to known thermophilic viruses reveals unexpected patterns of distribution and prevalence of viral sequences. PSV, the virus with the most extensive sequence similarity, was isolated from Obsidian hot spring (74°C, pH 5.6), about 30 km away from both Octopus and Bear Paw. The geochemistry of this thermal feature is distinct from the springs in this study (72) and life within includes a highly diverse population of *Archaea* and *Bacteria* (7, 39), most of which have not been detected in Octopus hot spring (10, 64) or elsewhere. Conversely, *Thermoproteus tenax spherical virus*, which is most similar to PSV in terms of sequence, morphology, and habitat (1), had very limited similarity to the YNP viral metagenomic sequences (not shown). The one other virus cultivated from Yellowstone, SSV-RH (85), had no significant tBLASTx similarity to any of the metagenomic samples. Other viruses showing high similarity to the metagenomic sequences were isolated on different continents, and, with the exception of YS40, occurred in highly acidic springs. This observation is remarkable because the microbial populations of acidic and neutral hot springs are quite distinct (66). In general, the viral populations described in this study seem disconnected from the microbial populations in the pools. Based on the sediments and filaments, Octopus spring is dominated by *Bacteria* (10, 63), which seems inconsistent with the large number of archaeal viruses seen in this study. Furthermore, the extensive conservation of viral sequences between the two hot springs in this study is surprising, given that microbial populations are highly temperature dependent (66) and the surface temperatures of these hot springs differ by 19°C (74°C vs. 93°C). The viral populations also appear much more diverse than would be predicted based on the diversity of microbes in the sediments and filaments (10, 63). A reasonable explanation for this disparity is that at least a portion of the viruses proliferate in the subterranean vent.

The regional and global conservation of viral sequences is an intriguing area for further study. There are examples of globally distributed genes among marine viral assemblages (12, 71). Since the oceans are contiguous across the earth, an obvious distribution mechanism exists. Groups of highly similar *Sulfolobus* viruses (85) and *Thermus* phages (88) have been isolated from thermal springs on different continents. In these cases, viruses were isolated from environments of similar pH and temperature and were cultivated on the same host under similar laboratory conditions. None of these selective conditions influenced this study, yet gene homologs to these viruses were detected. Conversely, most crenarchaeal virus morphotypes have been detected in enrichments from YNP (63, 67); however, little is known about conservation of genes in these enrichments.

The mechanism and basis of this conservation of viral sequence is open to speculation. It is possible that viruses sharing common genes adapt to the different host populations of the environment. Alternatively, hot springs may be inoculated by airborne viruses from other springs. It is also possible that the viruses acquire sequences from mesophilic viruses, although this explanation has no support in this study. Lineages of conserved viral genes may be older than the separation of the continents. Another explanation is proliferation of the viruses deeper in the vent. Thermophilic *Bacteria* and *Archaea*, potential hosts for viruses, have been detected in thermal aquifers several km beneath the earth's surface at abundances similar to those measured in this study (49) and many are distributed worldwide. While, it is impossible to separate the contribution of the subsurface viruses from any proliferation at the surface in the two pools in this study, samples from thermal springs with no pool at all, colleted within seconds of their emergence, have similar or somewhat higher viral abundances to those measured in this report (16). The low viral abundances and virus:microbe ratios (Table 1) relative to moderate temperature environments (86) also argue against dramatic proliferation of viruses at the surface. If a significant portion of the viruses in this study proliferate in the subsurface, the habitable portion of the subterranean aquifer could be continuous across much of the Yellowstone caldera or even much larger areas. A second implication is that, given the higher pressures in the vents, the temperature limit of life in the subterrestrial aquifers could greatly exceed the temperatures measured at the surface.

ACKNOWLEDGEMENTS

We thank our associates at Lucigen and MSU, especially Ronald Godiska for critical reading, Alice Ortmann for critical reading and epifluorescence microscopy and Sue Brumfield for electron microscopy. We thank Kendra Mitchel, Ann Rodman, Carrie Guiles, Christie Hendrix and John Varley for help with permitting and site identification, Forest Rohwer and Mya Breitbart for methods development, David Prangishvili for sharing data, and Jaysheel Bhavsar and Kanika Thapar for help with tBLASTX analysis. We also acknowledge the contributions of several anonymous reviewers. These samples were collected under Research Permit YELL-05240. This work was supported by NSF Grants 0109756 and 0215988 and NIH-NHGRI grant 1 R43 HG002714-01 to TS and DOE DE-FG02492 - 493 02ER83484 to DAM and the Delaware NSF EpSCOR program.

REFERENCES

1. **Ahn, D.G., S. I. Kim, J. K. Rhee, K. P. Kim, J. G. Pan and J. W. Oh.** 2006. TTSV1, a new virus-like particle isolated from the hyperthermophilic crenarchaeote Thermoproteus tenax. *Virology*. **351**, 280-290.
2. **Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new eneration of protein database search programs. Nucleic Acids Res. **25,** 3389-3402.
3. **Altschul, S.F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215**, 403-410
4. **Angly FE, B. Felts, M. Breitbart, P. Salamon, R. A. Edwards, C. Carlson, A. M. Chan, M. Haynes, S. Kelley, H. Liu, J. M. Mahaffy, J. E. Mueller, J. Nulton, R. Olson, R. Parsons, S. Rayhawk, C. A. Suttle, and F. Rohwer.** 2006. The marine viromes of four oceanic regions. PLoS Biol. **4(11)**:e368.
5. **Angly, F., B. Rodriguez-Brito, D. Bangor, P. McNairnie, M. Breitbart, P. Salamon, B. Felts J. Nulton, J. Mahaffy and F. Rohwer.** 2005. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics. **6:** 41.
6. **Arnold, H.P., W. Zillig, U. Ziese, I. Holz, M. Crosby, T. Utterback, J. F. Weidmann, J. K. Kristjanson, H. P. Klenk, K. E. Nelson and C. M. Fraser**. 2000. A novel lipothrixvirus, SIFV, of the extremely thermophilic crenarchaeon Sulfolobus. Virology. **267:** 252-266.
7. **Barns, S.M., R. E Fundyga, M. W. Jeffries and N. R. Pace**. 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc. Natl. Acad. Sci. USA. **91:** 1609-1613.
8. **Barry, E. R. and S. D. Bell.** 2006. DNA replication in the archaea. Microbiol. Mol. Biol. Rev517 . **70(4)**:876-887.
9. **Bench, S. R., T. E. Hanson, K. E. Williamson, D. Ghosh, M. Radosovich, K. Wang, K. E. Wommack.** 2007. Metagenomic Characterization of Chesapeake Bay Virioplankton. Appl Environ Microbiol. in press.
10. **Blank, CE, S. L. Cady and N. R. Pace**. 2002. Microbial composition of near-boiling silica depositing thermal springs throughout Yellowstone National Park. Appl. Environ. Microbiol. **68(10)**:5123-5135.
11. **Blum, H, W. Zillig, S. Mallok, H. Domdey and D. Prangishvili.** 2001. The genome of the archaeal virus SIRV1 has features in common with genomes of eukaryal viruses. Virology. **281(1)**:6-9.
12. **Breitbart, M. and F. Rohwer.** 2005. Here a virus, there a virus, everywhere the same virus? Trends Microbiol. **6:**278-284.
13. **Breitbart, M., B. Felts, S. Kelley, J. M. Mahaffy, J. Nulton, P. Salamon and F. Rohwer.** 2004. Diversity and population structure of a near-shore marine-sediment viral community. Proc. Biol. Sci. **271**: 565-574.
14. **Breitbart, M., I. Hewson, B. Felts, J. M. Mahaffy, J. Nulton, P. Salamon and F. Rohwer** 2003. Metagenomic analyses of an uncultured viral community from human feces. J. Bact. **85:** 6220-6223.
15. **Breitbart, M., P. Salamon, B. Andresen, J. M. Mahaffy, A. M. Segall, D. Mead, F. Azam, and F. Rohwer.** 2002. Genomic analysis of uncultured marine viral communities. Proc. Natl. Acad. Sci. USA. **99:**14250-14255.
16. **Breitbart, M., L. Wegley, S. Leeds, T. Schoenfeld and F. Rohwer.** 2004. Phage community dynamics in hot springs. Appl. Environ. Microbiol. **70:**1633-1640.
17. **Brock, T.D.** 1978. Thermophilic microorganisms and life at high temperatures New York: Springer-Verlag.
18. **Brown, J. R., W. F. Doolittle.** 1997. Archaea and the prokaryote-to-eukaryote transiti543 on. Microbiol Mol Biol Rev. **61**(4):456-502.

19. **Büchen-Osmond C.** 2003. Taxonomy and Classification of Viruses. In: Manual of Clinical Microbiology, 8th ed, Vol 2, p. 1217-1226, ASM Press, Washington DC.
20. **Canchaya, C., G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann and Brussow, H.** 2003. Phage as agents of lateral gene transfer. Curr. Opin. Microbiol. **6:**417-424.
21. **Cann, A. J., S. E. Fandrich, and S. Heaphy.** 2005. Analysis of the virus population present in equine faeces indicates the presence of hundreds of uncharacterized virus genomes. Virus Genes. **30:** 151-156.
22. **Coenye, T. and P. Vandamme.** 2003. Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol. Lett. **228:**45-49.
23. **Culley A. I., A. S. Lang and C. A. Suttle.** 2007. The complete genomes of three viruses assembled from shotgun libraries of marine RNA virus communities. Virol J. **6**:4:69.
24. **Daubin V. and H. Ochman.** 2004. Start-up entities in the origin of new genes. Curr. Opin. Genet. Dev. **14(6)**:616-619.
25. **Dionne I., and S. D. Bell.** 2005. Characterization of an archaeal family 4 uracil DNA glycosylase and its interaction with PCNA and chromatin proteins. Biochem. J. **387**:859-863.
26. **Feng, L, W. Wang, J. Cheng, Y. Ren, G. Zhao, C. Gao, Y. Tang, X. Liu, W. Han, X. Peng, R. Liu and L. Wang.** 2007. Genome and proteome of long-chain alkane degrading Geobacillus thermodenitrificans NG80-2 isolated from a deep-subsurface oil reservoir. Proc. Natl. Acad. Sci. USA. **104**: 5602-5607.
27. **Filee J., P. Forterre, T. Sen-Lin and J. Laurent.** 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J. Mol. Evol. **54(6)**:763-73.
28. **Filee J., P. Forterre, and J. Laurent.** 2003. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. Res. Microbiol. **154:**237-243.
29. **Forterre P.** 2006. The origin of viruses and their possible roles in major evolutionar569 y transitions. Virus Res. **117:**5-16.
30. **Forterre P.** 2005. The two ages of the RNA world, and the transition to the DNA world: a story of viruses and cells. Biochimie. **87(9-10)**:793-803.
31. **Fournier, R.O.** 2005. Geochemistry and dynamics of the Yellowstone National Park Hydrothermal System, 3-30. *In* W. P., Inskeep and T. R. McDermott (ed.) Geothermal Biology and Geochemistry in YNP. Thermal Biology Institute, Bozeman, MT.
32. **Geslin, C., M., Le Romancer, G. Erauso, M. Gaillard, G. Perrot and D. Prieur**. 2003. PAV1, the first virus-like particle isolated from a hyperthermophilic euryarchaeote, "Pyrococcus abyssi". J. Bacteriol. **185:**3888-3894.
33. **Godiska, R., M. Patterson, T. Schoenfeld and D. Mead.** 2005. Beyond pUC: Vectors for Cloning Unstable DNA. in *DNA Sequencing: Optimizing the Process and Analysis* ed. Kieleczawa, J. (Jones and Bartlett) pp. 55-75.
34. **Gold, T.** 1992. The deep, hot biosphere. Proc. Natl. Acad. Sci. USA. **89:**6045-6049.
35. **Haft, D. H., J. Selengut, E. F. Mongodin and K. E. Nelson.** 2005. A guild of 45 CRISPR associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol. **1(6)**:e60.
36. **Haring M, X. Peng, K. Brugger, R. Rachel, K. O. Stetter, R. A. Garrett and D. Prangishvili.** 2004. Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera Pyrobaculum and thermoproteus: a novel virus family, the Globuloviridae. Virology. **323(2)**:233-242.
37. **Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, Ford ME, Gonda RM, Houtz JM, Hryckowian AJ, Kelchner VA, et al.** 2006. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet. **2(6)**: e2
38. **Hjörleifsdottir, S. H., G. O. Hreggvidsson, O. H. Fridjonsson, A. Aevarsson and J. K. Kristjansson.** (December 10, 2002) U. S. Patent 6,492,161.
39. **Hugenholtz, P., C. Pitulle, K. L. Hershberger and Pace, NR.** 1998. Novel division level bacterial diversity in a Yellowstone hot spring. J. Bacteriol. **180**, 366-376.
40. **Jahnke, L. L., W. Eder, R. Huber, J. M. Hope, K.-I. Hinrichs, J. M. Hayes, D. J. Des Marais, S. L. Cady, and R. E. Summons.** 2001. Signature lipids and stable carbon isotope analyses of Octopus Spring hyperthermophilic communities compared with those of Aquificales representatives. Appl. Environ. Microbiol. **67**:5179–5189.
41. **Kahler, M. and G. Antranikian.** 2000. Cloning and characterization of a family B DNA polymerase from the hyperthermophilic crenarchaeon Pyrobaculum islandicum. J Bacteriol. **182(3)**:655-63.
42. **Kessler, A., A. B. Brinkman, J. van der Oost J and D. Prangishvili.** 2004 Transcription of the rod-shaped viruses SIRV1 and SIRV2 of the hyperthermophilic archaeon sulfolobus. J Bacteriol. **186(22)**:7745-53.
43. **Kwan, T., J. Liu, M. DuBow, P. Gros and J.Pelletier.** 2005. The complete genomes and proteomes of 27 Staphylococcus aureus bacteriophages. Proc. Natl. Acad. Sci. USA. **102**, 5174- 5179.
44. **Lindell, D., M. B. Sullivan, Z. I. Johnson, A. C. Tolonen, F. Rohwer and S. W. Chisholm.** 2004. Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc. Natl. Acad. Sci. USA.**101:**11013-11018.
45. **Lukashin, A. and M. Borodovsky.** 1998. GeneMark.hmm: new solutions for gene finding. Nucl. Acids, Res. **26(4)**:1107-1115.
46. **Lucchini, S., F. Desiere and H. Brussow.** 1998. Comparative genomics of Streptococcus thermophilus phage species supports a modular evolution theory. *Virology.* **246**, 63-73.
47. **Martin A., S. Yeats, D. Janekovic, W. D. Reiter, W. Aicher and W. Zillig.** 1984. SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12. EMBO J. **3(9)**:2165-2168.
48. **McCleskey, R.B., J. W. Ball, D. K. Nordstrom, J. M. Holloway and H. E. Taylor.** 2004, Water-Chemistry Data for Selected Hot Springs, Geysers, and Streams in Yellowstone National Park, Wyoming, 2001-2002. U.S. Geological Survey Open-File Report 2004-1316.
49. **Moser, D.P., T. M. Gihring, F. J. Brockman, J. K. Fredrickson, D. L. Balkwill, M. E. Dollhopf, B. S. Lollar, L. M. Pratt, E. Boice, G. Southam, *et al*.** 2005. Desulfotomaculum and Methanobacterium spp. dominate a 4- to 5-kilometer-deep fault. Appl. Environ. Microbiol. **71**:8773-8783.
50. **Naryshkina, T., J. Liu, L. Florens, S. K. Swanson, A. R. Pavlov, N. V. Pavlova, R. Inman, L. Minakhin, S. A. Kozyavkin and M. Washburn, *et al.*** 2006. Thermus thermophilus bacteriophage phiYS40 genome and proteomic characterization of virions. J. Mol. Biol. **364:**667-677.
51. **Noble, R.T. and J. A. Fuhrman**. 1998. Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquat. Microb. Ecol. **14**:113–118.
52. **Paul, J.H., S. J. Williamson, A. Long, D. John, A, Segall, and F. Rohwer.** 2005. Complete genome sequence of phiHSIC, a pseudotemperate marine phage of Listonella pelagia. Appl. Environ. Microbiol. **71:**3311-3320.
53. **Pedersen, K.** 2000. Exploration of deep intraterrestrial life – current perspectives. FEMS Microbiology Letters. **185:** 9-16.
54. **Pedulla, M.L., M. E. Ford, J. M. Houtz, T. Karthikeyan, C. Wadsworth, J. A. Lewis, D. Jacobs-Sera, J. Falbo, J. Gross, N. R. Pannunzio, *et al.*** 2003. Origins of highly mosaic mycobacteriophage genomes. Cell**. 113:**171-182.
55. **Peng X., H. Blum, Q. She, S. Mallok, K. Brugger, R. A. Garrett, W. Zillig and D. Prangishvili.** 2001. Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. Virology. **291(2)**:226-234.
56. **Peng X., A. Kessler, H. Phan, R. A. Garrett and D. Prangishvili.** 2004. Multiple variants 645 of the archaeal DNA rudivirus SIRV1 in a single host and a novel mechanism of genomic variation. Mol Microbiol. **54(2)**:366-375.
57. **Prangishvili D., R. A. Garrett and E. V. Koonin.** 2006. Evolutionary genomics of archaeal viruses: unique viral genomes in the third domain of life. Virus Res. **117(1)**:52-67.
58. **Prangishvili D. and R. A. Garrett.** 2004. Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. Biochem Soc Trans. **32(2)**:204-208.
59. **Prangishvili D., K. Stedman and W. Zillig.** 2001. Viruses of the extremely thermophilic archaeon Sulfolobus. Trends Microbiol. **9(1)**:39-43.
60. **Prangishvili D.** 2003. Evolutionary insights from studies on viruses of hyperthermophilic archaea. Res. Microbiol. **154(4)**:289-294.
61. **Prangishvili, D. and R. A. Garrett.** 2005. Viruses of hyperthermophilic Crenarchaea. Trends Microbiol. **13:**535-542.
62. **Prangishvili, D., G. Vestergaard, M. Haring, R. Aramayo, T. Basta, R. Rachel and R. A. Garrett.** 2006. Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle. J. Mol. Biol. **359:**1203-1216.
63. **Rachel, R., M. Bettstetter, B. P. Hedlund, M. Haring, A. Kessler, K. O. Stetter and D. Prangishvili.** 2002. Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. Arch. Virol. **147:**2419-2429.
64. **Reysenbach A.L., G. S. Wickham and N. R. Pace.** 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. Appl. Environ. Microbiol. **60:**2113-2119.
65. **Reysenbach, A.L. and E. Shock.** 2002. Merging genomes with geochemistry in hydrothermal ecosystems. Science. **296:**1077-1082.
66. **Reysenbach, A.L., D. Gotz and D. Yernool.** 2002. Microbial Diversity of Marine and Terrestrial Thermal Springs. *In* J. T Staley and A.-L.Reysenbach (ed.) Biodiversity of Microbial Life. Wiley Liss, New York.
67. **Rice, G., K. Stedman, J. Snyder, B. Wiedenheft, D. Willits, S. Brumfield, T. McDermott and M. J. Young.** 2001. Viruses from extreme thermal environments. Proc. Natl. Acad. Sci. USA. **98:**13341-13345.

68. **Roberts, J. A., S. D. Bell and M. F. White.** 2003. An archaeal XPF repair endonuclease dependent on a heterotrimeric PCNA. Mol Microbiol. **48(2)**:361-371.
69. **Sakaki Y. and T. Oshima.** 1975. Isolation and characterization of a bacteriophage infectious to an extreme thermophile, Thermus thermophilus HB8. J. Virol. **15(6)**:1449-1453.
70. **Seguritan, V., I. Feng, F. Rohwer, M. Swift and A. M. Segall.** 2003. Genome sequences of two closely related Vibrio parahaemolyticus phages, VP16T and VP16C. J. Bact. **185**:6434- 6447.
71. **Short, C. M. and C. A. Suttle.** 2005. Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl Environ Microbiol. **71(1)**:480-6.
72. **Shock, E.L., M. Holland, D. R. Meyer-Dombard and J. P. Amend.** 2005. Geochemical sources of energy for microbial metabolism in hydrothermal ecosystems: Obsidian Pool, Yellowstone National Park. p. 95-112. *In* W. P., Inskeep and T. R. McDermott (ed.) Geothermal Biology and Geochemistry in YNP. Thermal Biology Institute, Bozeman, MT.
73. **Snyder J. C., J. Spuhler, B. Wiedenheft, F. F Roberto, T. Douglas and M. J. Young.** 2004. Effects of culturing on the population structure of a hyperthermophilic virus. Microb Ecol. **48(4)**:561-6.
74. **Snyder, J. C., K. Stedman, G. Rice, B. Wiedenheft, J. Spuhler and M. J. Young.** 2003. Viruses of Hyperthermophilic Archaea. Res Microbiol. **154(7)**:474-82.
75. **Stoner, D.L., M. C. Geary, L. J. White, R. D. Lee, J. A. Brizzee, A. C. Rodman and R. C. Rope.** 2001.Mapping microbial biodiversity. Appl. Environ. Microbiol. **67:**4324-4328.
76. **Sullivan, M.B., M. Coleman, P. Weigele, F. Rohwer and S. W. Chisholm. 2005.** Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS Biol. **3:**1-17.
77. **Suttle C. A.** 2007. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol. **5(10)**:801-12.
78. **Tatusov, R. L., E. V. Koonin and D. J. Lipman.** 1997. A genomic perspective on protein families. Science. **278**:631-637.
79. **Vestergaard G., M. Haring, X. Peng, R. Rachel, R. A. Garrett and D. Prangishvili.** 2005. A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus Acidianus. Virology. **336(1)**:83-92.
80. **Villarreal, L. P. and V. R. DeFilippis.** 2000. A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. J. Virol. **74:**7079-7084.
81. **Wang, I. N., D. L. Smith and R. Young**. 2000. Holins: the protein clocks of bacteriophage infections. Annu. Rev. Microbiol. **54:**799-825.
82. **Ward, D. M., M. J. Ferris, S. C. Nold and M. M. Bateson.** 1998. A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. Microbiol Mol Biol Rev. **62(4)**:1353-70.
83. **Weinbauer, M. G. and F. Rassoulzadegan.** 2004. Are viruses driving microbial diversification and diversity? Environ Microbiol. **6(1)**:1-11.
84. **Wen, K., A. C. Ortmann and C. A. Suttle.** 2004. Accurate estimation of viral abundance by epifluorescence microscopy. Appl Environ Microbiol. **70(7)**:3862-3867.
85. **Wiedenheft, B., K. Stedman, F. Roberto, D. Willits, A. K. Gleske, L. Zoeller, J. Snyder, T. Douglas and M. Young.** 2004. Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses. J. Virol. **78:**1954-1961.
86. **Wommack, K.E. and R. R. Colwell.** 2000. Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. **64:**69-114.
87. **Xiang, X., L. Chen, X. Huang, Y. Luo, Q. She and L. Huang.** 2005. Sulfolobus tengchongensis spindle-shaped virus STSV1: virus-host interactions and genomic features. J. Virol. **79:**8677-8686.
88. 88. **Yu, M.X., M. R. Slater and H. W. Ackermann.** 2006. Isolation and characterization of Thermus bacteriophages. Arch. Virol. **151:**663-679.