APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 2008, p. 4164–4174 Vol. 74, No. 13
0099-2240/08/$08.00+0 doi:10.1128/AEM.02598-07

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, 2120 W. Greenview Drive, Suite 9, Middleton, Wisconsin 53562[1]; U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 94598[2]; Department of Plant and Soil Sciences, University of Delaware, 15 Innovation Way, Newark, Delaware 19711[3]; and Plant Sciences and Plant Pathology, Montana State University, 307 AgBio Building, Bozeman, Montana 59717[4]

Received 16 November 2007/Accepted 16 April 2008

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 the assembly identity reduced the estimated number of viral types from 1,440 and 1,310 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, nonlytic infection lifestyle, analysis of coding sequences suggests lytic viruses are more common in geothermal environments than previously thought. The 50% assembly included one contig with 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 showed 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.

Subterranean 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 74 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 (53, 65). 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 to 270°C are found at depths of 100 to 550 m 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 10^{30} viruses in the world's oceans (77) may comprise several hundred thousand different species (5). 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 (5, 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 those 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 to 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 (47, 69), hundreds of bacteriophages (88), dozens of crenarchaeal viruses (reviewed in references 59, 61, and 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*, *Tec-*

* Corresponding author. Mailing address: Lucigen Corporation, 2120 West Greenview Drive, Suite 9, Middleton, WI 53562. Phone: (608) 831-9011, ext. 222. Fax: (608) 831-9012. E-mail: tschoenfeld@lucigen.com.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 25 April 2008.

TABLE 1. Sample sites and viral and microbial counts

Hot spring	Temp (°C)	pH	No. of cells/ml	No. of viruses/ml	Virus/microbe ratio	No. of viruses/ml in concentrate	Theoretical[a] no. of viruses/ml	Efficiency (%)
Bear Paw	74	7.34	4.3×10^6	1.44×10^6	0.33	1.48×10^8	7.21×10^9	2.1
Octopus	93	8.14	9.0×10^5	3.07×10^5	0.34	2.18×10^8	1.53×10^9	14.2

[a] Based on a concentration factor of 5,000× (500 liters to 100 ml).

tiviridae, and *Inoviridae* (88). Their morphologies and the available genomic sequences (38, 50) 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 contents suggest that 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 (56). 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., rRNA genes) 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 the composition of viral assemblages through studies of gene homology. No metagenomic analysis of waterborne viral populations in geothermal environments has been reported. In fact, planktonic life in thermal environments is underexplored 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) (44°33′21.994′′N, 110°50′5.232′′W) and Octopus (44°32′Z.701′′N, 110°47′52.402′′W) hot springs (Table 1) (75). The temperatures of the hot springs are based on direct measurement on the day of sampling. The pH values were determined by the U.S. Geological Survey (48). Additional geochemical data for Octopus hot spring is available in Table S1 in the supplemental material. Thermal water (400 to 600 liters) was filtered using a 100,000-molecular cutoff (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-kDa filter. Viral concentrates were imaged by transmission electron microscopy (Leo 912AB operating at 80 kV). 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 h before being counted. Immediate 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 12,000 rpm for 20 min, syringe filtered using a 0.2-μm Acrodisc filter (Gelman), and further concentrated to 400 μl by filtration using a 30,000-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 $MgSO_4$, 50 mM Tris-HCl, pH 7.5) using a 30,000-MWCO spin filter. Benzonase endonuclease (Sigma; 10 U) was added, and the reaction mixtures were incubated for 30 min at 23°C. EDTA (20 mM), sodium dodecyl sulfate (0.5%), and proteinase K (100 U) were added, and the reaction mixtures were incubated for 3 h at 56°C. NaCl (0.7 M) and cetyltrimethylammonium bromide (1%) were added, and DNA was extracted with phenol-chloroform and ethanol precipitated.

Library construction and sequencing. Viral DNA was physically sheared to 3 to 6 kb using a HydroShear device (Genomic Solutions, Michigan). The ends were made blunt using the DNATerminator end repair kit (Lucigen, Middleton, WI), and the fragments were ligated to a double-stranded asymmetrical linker comprised of one blunt phosphorylated end (5′-GATGCGGCCGCTTGTATC TGATACTGCT-3′; linker 1) and one nonphosphorylated, staggered end (5′-G GAGCAGTATCAGATACAAGCGGCCGCATC-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- to 6-kb fragments. These fragments were PCR amplified using Vent DNA polymerase (New England Biolabs, Massachusetts) and a primer targeted to linker 1 (5′-AGCAGTATCAGATACAA GCGGCCGCATC-3′). The amplification products were gel purified again, inserted into the cloning site of the transcription-free pSMART vector (Lucigen), and used to transform *Escherichia coli* 10G cells (Lucigen). Libraries were sequenced by the Department of Energy 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 protein database (GenBank) using BLASTx (2, 3). The 50 most significant BLASTx scores ($E < 10^{-3}$) were recorded. The occurrences of key words in the output of the BLASTx search were counted using PERL scripts written for this project, and the sequences were categorized by function. The sequences were assembled using the SeqMan program (DNAStar, Wisconsin) at a minimum of 50% or 95% identity over a minimum of 20 nucleotides. Metagenome sequence libraries were compared to each other and to all the sequences in GenBank using tBLASTx (NCBI) with a cutoff E value of $<10^{-3}$. Where indicated, the apparent open reading frames (ORFs) were identified and translated using the GeneMark program (46). These translations were compared to the nonredundant protein database using the BLASTp program. The rank abundances were calculated using PHAge Community from the Contig Spectrum (PHACCS) web utility located at http://phage.sdsu.edu/research/tools/phaccs/ (4) based on an average genome length of 50 kb.

RESULTS AND DISCUSSION

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. Although the pH values of these hot springs are both circumneutral, the temperatures and apparent microflora differ widely. Bear Paw is significantly cooler and is characterized by orange sedentary microbial growth in the pool. Octopus water emerges at the boiling point at the local elevation of 2,300 m, with none of the orange growth. Octopus hot spring is well documented to support prolific microbial life (17), and its geochemistry (48) (see Table S1 in the supplemental material) is suitable for chemotrophic lifestyles. Reported analyses based

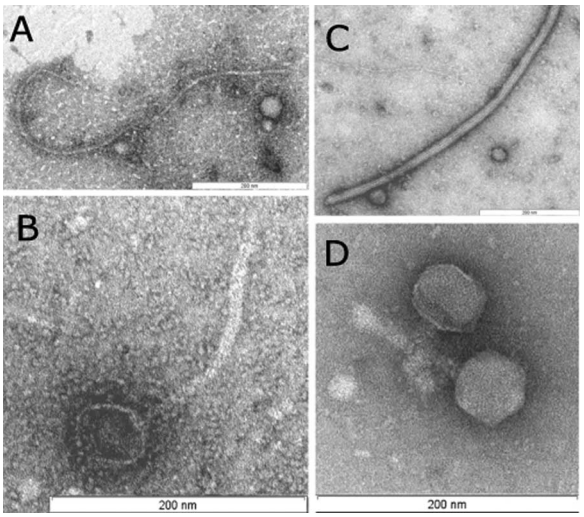

FIG. 1. Transmission electron microscopy images of virus-like particles directly isolated from YNP hot springs. Images from Bear Paw (A and B) and Octopus (C and D) hot springs are shown. The bar in each panel is 200 nm.

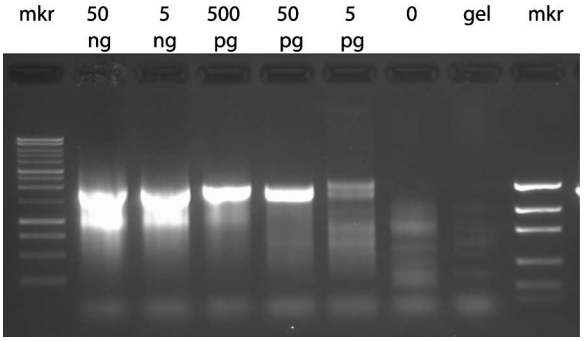

FIG. 2. Sensitivity of linker-facilitated anonymous-DNA amplification. Decreasing amounts of lambda DNA (50 ng, 5 ng, 500 pg, 50 pg, and 5 pg, as indicated) were sheared, end repaired, linker ligated, and size selected on an agarose gel. The resulting material was amplified using Vent DNA polymerase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

on rRNA gene sequences (10, 64) show that microbial diversity is relatively limited compared to moderate-temperature environments. These studies and studies of lipid and isotope compositions (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. To our knowledge, no detailed study of the chemical composition or life in Bear Paw has been published.

Virus-enriched fractions were isolated from 400 to 600 liters of hot-spring water for library construction and sequence analysis. Viral abundances (Table 1) were at the lower end of the range of 10^4 to 10^9 reported for thermal springs in California (16) and moderate-temperature aquatic environments (86). The virus/microbe ratios in the hot springs were much lower than in moderate-temperature environments (typically 3 to 10). These low virus/microbe ratios 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 the microbial population is reduced. Actual yields of viruses were significantly below theoretical yields (Table 1) for both hot springs. 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 (Fig. 1). The morphologies of viral particles in the concentrates represented most morphological families of known thermophilic viruses. 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 ag of DNA per virus, abundances of 10^5 to 10^6 viruses per ml correspond to 5 to 50 ng of viral DNA per liter. In practice, processing of hundreds of liters of spring water yielded at most 100 ng of DNA, much less than is normally required for library construction. This low yield of virus precluded cesium chloride purification of the viral particles, as is commonly done 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- to 8-kb insert libraries with none of the potential modified nucleotides. This library construction method has been used in the analysis of several cultivated and uncultivated viral genomes (9, 13, 14, 15, 44, 52, 70, 76) but has never been fully described. It is described in detail here for the first time. Viral DNA was physically sheared, and short (20-bp) 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 provided identical primer annealing sites for each viral template in the mixture, which significantly limited amplification bias. An important limitation of this approach is that it selects exclusively for double-stranded DNA viruses. All cultivated thermophilic bacteriophage and archaeal viruses have double-stranded DNA genomes, except certain *Thermus*-specific inoviruses, which have single-stranded DNA 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.

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 (Fig. 2). Comparison of 50 clones to lambda whole-genome sequence showed 22 discrepancies among 39,000 bases sequenced, an error rate of 0.05%. Coverage was 64.5% with no apparent

TABLE 2. Functional grouping of predicted genes in the viral metagenomes

COG functional category	No. of reads matching a keyword		% with a keyword match	
	Bear Paw	Octopus	Bear Paw	Octopus
No BLASTx similarity	2,545	8,469		
F (nucleotide transport and metabolism)	1,445	2,130	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	1,045	23.19	18.55

stacking of reads. While this level of coverage does not preclude cloning bias or sequence dropout, assembly lengths had a normal distribution, suggesting that any cloning bias was minimal, and several sequences known to be toxic (e.g., genes for holins and lysozymes) were among the clones in the library.

A total of 28,883 sequence reads were determined from Bear Paw (7,685 reads) and Octopus (21,198 reads) hot springs. Paired-end reads averaged 981 nucleotides each, or nearly 30 Mb in 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 represented 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 those of the Octopus library, as determined by several cases of sequence stacking.

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 rRNA gene 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* species that grow at the temperature of Bear Paw contain up to 30 rRNA genes (26). BLASTn analysis identified only four rRNA gene 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 that viral enrichment was quite high, particularly for the Octopus library. This inference is supported by 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; see Table S2 in the supplemental material), is consistent only with a predominantly viral origin of the sequences.

Identification of likely gene products and viral lifestyles. BLASTx analysis of the individual reads was used to identify coding sequences in the libraries. While most reads revealed no significant similarity to known proteins (i.e., no BLASTx similarity) (Table 2), a significant portion of the sequences could be assigned an apparent function based on BLASTx analysis (see Table S2 in the supplemental material). The majority of these predicted functions fall into 5 of the 23 NCBI cluster of orthologous groups (COG) functional categories (78) or are virus-specific functions that have no assigned COG function, e.g., lysin, packaging, capsid, tail, or tape measure protein (Table 2). These five COG categories are all nucleic acid metabolism, information-processing, and translation-related functions, which are commonly associated with phage and viruses.

Certain similarities were particularly interesting. The 532 lysin-like genes among 600 viral equivalents suggest that lytic viruses are quite common in the hot springs, in contrast to the cultured thermophilic crenarchaeal viruses, all of which are nonlytic. 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). 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 (*Acidianus two-tailed virus* [ATV], *Sulfolobus tengchongensis spindle-shaped virus* 1 [STSV1], and four *Sulfolobus spindle-shaped virus* isolates) (62, 85, 87) and induction of prophage by mitomycin C in 1 to 9% of hot-spring microbial cells (16).

Viruses and lateral gene transfer in thermal environments. 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 references 29 and 60). Gene similarities seen in the metagenomic libraries (see Table S2 in the supplemental material) 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

TABLE 3. Sequence assembly data and estimation of viral diversity

Parameter	Value at assembly identity[a]:			
	95%		50%	
	Bear Paw	Octopus	Bear Paw	Octopus
No. of contigs assembled	6,191	13,543	4,850	4,788
Avg. no. of reads per contig	1.239	3.129	1.587	4.427
Largest contig (nucelotides)	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-Wiener score	6.88	6.85	5.88	5.29

[a] Numbers of sequence reads were as follows: Bear Paw, 7,685; Octopus, 21,198; total, 28,883.

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

Virus	Reference(s)	Accession no.	No. of tBLASTx similarities	
			Bear Paw	Octopus
ARV	79	AJ875026	36	228
SIRV1 and -2	11, 55	AJ344259, AJ414696	30	217
PSV	36	AJ635161	44	152
SIFV	6	AF440571	7	46
S. tengchongensis spindle-shaped virus 1	87	AJ783769	26	22
ATV	62	AJ888457	8	17
YS40	50	DQ997624	15	41
Thermoproteus tenax spherical virus 1	1	AY722806	6	12
Twort	43	AY954970	4	21

examples of nonorthologous replacement of cellular genes by viral genes (28). 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).

Also common in the metagenomic libraries are presumptive ribonucleotide reductases (14 and 50, respectively) and thymidylate synthase (7 and 51, respectively) genes. The conservation of these genes between 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 200 apparent *pol* gene homologs were identified in the two metagenomic libraries, with all the polymerase families represented. 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). The high abundance of both *pol* and *lys* 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.

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 >95% identity over 20 nucleotides as the assembly stringency (5, 13, 14, 15). Using this criterion, the power law rank-abundance model built into the PHACCS tool (4) predicted 1,400 and 1,310 viral types in the Bear Paw and Octopus hot springs, respectively, with no one viral type representing more than about 2% of the population (Table 3). For reference, 1,650, 3,350, 7,180, 7,340, and 2,390 viral genotypes were reported in estuarine, near-shore marine, open-ocean, marine sediment, and fecal-viral assemblages, respectively (5, 9, 13, 14, 15), with no single viral species representing more than 2 to 3% in any case.

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 (see Fig. 4) (45). Genetic diversity outside the conserved regions is probably higher than these models indicate. Second, the generation of new viral species by mosaicism, modular evolution, or lateral gene transfer (20, 80, 83) would not be detected using an 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. Finally, 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 classical criteria (host range, morphology, replication lineages, and physicochemical and antigenic properties) (19) are considered to be related (37, 43, 45), although they may share as little as 50% nucleotide identity over much of their genomes.

Lower-stringency assemblies reveal population heterogeneity. To accommodate the genomic heterogeneity inherent in viral populations, sequences were also assembled at 50% identity (Table 3). As expected, the numbers of viral types decreased to 548 and 283 in Bear Paw and Octopus, respectively. These lower-stringency assemblies proved quite useful for associating sequences of related, but not identical, viral types and for studying diversity among these 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 >1,000 reads each (see Table S3 in the supplemental material). In each case, reads were evenly distributed across the contigs. The 17 >10-kb contigs comprise a total of 7.04 Mbp (33% of the total metagenomic sequence), or about 140 viral equivalents. The four strongest BLASTx hits to the 35-kb contig belonged to the thermophilic crenarchaeal viruses *Acidianus rod-shaped virus* (ARV), *Sulfolobus islandicus rod-shaped virus* 1 (SIRV1) and SIRV2, and *Sulfolobus islandicus filamentous virus* (SIFV) (Table 4). The only significant similarity for the 19-kb contig was to the thermophilic crenarchaeal virus *Pyrobaculum spherical virus* (PSV). 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

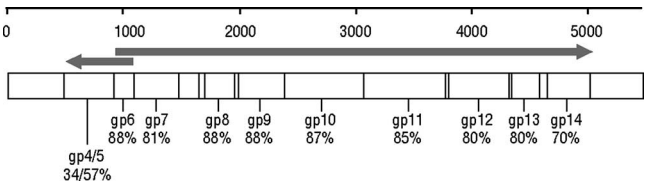

FIG. 3. Genes and gene order are highly conserved between a cultured crenarchaeal virus and a consensus contig from the Bear Paw library. Contig 372 (5,492 bp; 71 reads) was assembled at ≥50% identity from the Bear Paw library. ORFs identified by the GeneMark algorithm were compared by BLASTp to proteins in GenBank. Similarities to PSV proteins are shown with percent coding identity. The gene names are based on the annotation in GenBank and are named in order of their locations on the viral chromosome. The directions of transcription are indicated by the arrows.

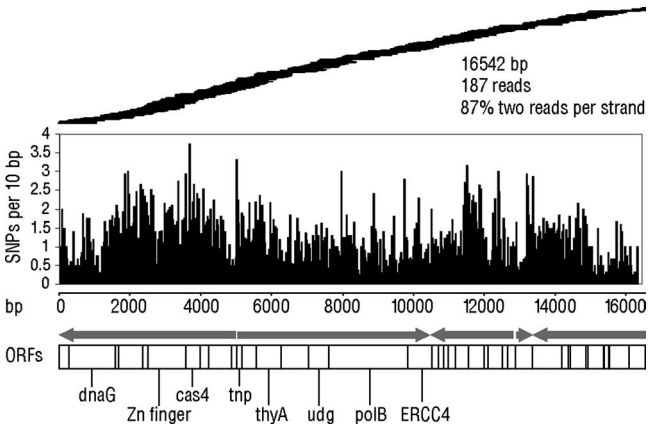

FIG. 4. Alignment of nucleotide polymorphisms with coding sequences in a 16.5-kb consensus contig from the Octopus hot spring. Contig 722 was assembled at ≥50% identity from the Octopus library. Sequence coverage is shown at the top, with each line representing a separate read. SNPs per 10 base pairs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

into 19 contigs of >4 kb. They included 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 (Fig. 3). Eleven ORFs were identified by the GeneMark algorithm (46). BLASTp analysis of each showed the strongest similarity to the putative coding sequences of PSV (36). Nucleotide identities were as high as 88%, gene order was perfectly preserved relative to the cultured virus, and gene overlaps were identical between the composite contig and the cultivated virus. Interestingly, two different ORFs of the PSV genome, gp4 and gp5, are apparently related to each other, since both had significant similarity to the same region of the consensus contig. In both the cultured viral 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 the population diversity of an assembled metagenome with the biochemistry of the gene products (Fig. 4). This 16.5-kb contig, assembled at 50% identity, includes 187 reads (average coverage, 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 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 had 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 *cas4* (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, single-nucleotide polymorphism (SNP) frequency was aligned with the 50% assembly consensus sequence of the contig. The 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 distributions in the *dnaG*, *udg*, *polB*, and ERCC homologs were 0.565, 0.617, 0.569, and 0.548 per 10 bp, respectively, while the distributions in the Zn finger, *cas4*, and *thyA* homologs were 0.979, 1.31, and 0.728, respectively. As additional biochemical and structural data become available for the replication proteins, molecular diversity can be correlated with variations in function and structure.

Similarities to known viral and microbial genomes imply phylogeny. tBLASTx analysis was used to infer the phylogenetic origins 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 (Fig. 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 four- to fivefold-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 (see above).

Interestingly, the libraries contained a sizable number of

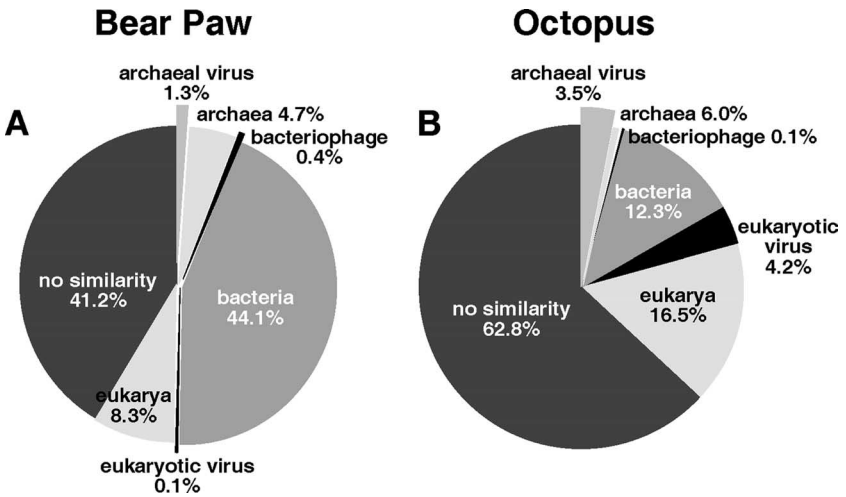

FIG. 5. Broad classification of viral metagenomic contigs based on tBLASTx similarities. Contigs assembled at 95% identity from Bear Paw and Octopus reads (A and B, respectively) were compared to sequences in GenBank to infer phylogeny. Shown are frequencies of contigs with no significant sequence similarity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

sequences with homology to eukaryotic genes, 16.5% for the Octopus spring and 8.3% for Bear Paw, which may reflect the commonly observed overlap in gene sequence homology between *Archaea* and *Eukarya* in general (18). Almost all known crenarchaeal viruses were cultivated on 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 the 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 order *Aquificales*.

Remarkably, the apparent phylogeny of viral populations in this study seems disconnected from the microbial populations in the pools, since the reported dominance of the Octopus spring by *Bacteria* (10, 63) appears inconsistent with the large number of archaeal viruses seen in the Octopus library. Furthermore, the viral populations also appear much more diverse than would be predicted based on the low diversity of microbes in the sediments and filaments.

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

Genus	No. of contigs	
	Bear Paw	Octopus
Archaea		
Pyrobaculum	124	684
Aeropyrum	62	626
Sulfolobus	38	326
Acidianus	25	185
Bacteria		
Aquifex	474	1,138

Alignment of the metagenome to cultivated viral genomes. Overall, only 3.4% of the high-stringency (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 nonthermophilic virus, phage Twort (43), was limited to the helicase gene, which shares similarity with that of SIFV (see above). The two libraries had similar frequencies of genetic homology to archaeal viruses and phage. Notable exceptions were ARV and SIRV1 and -2, where the Octopus library demonstrated a higher frequency of homology than the Bear Paw library, and the STSV1 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 (Fig. 6). Almost the entire genome of PSV has similarity to sequences in both metagenomic libraries, with median identities of 60% and 51% to Bear Paw and Octopus, respectively. Sequence similarities to the other crenarchaeal viruses and to *Thermus thermophilus* bacteriophage YS40 were limited to a few specific ORFs, but the degrees of similarity were 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 (Fig. 6) (1, 6, 11, 36, 42, 50, 55, 62, 79, 87; D. Prangishvili, personal communication). These regions of high conservation are genes associated with virion components, DNA replication, transposition, recombination, or nucleic acid metabolism.

The degree of alignment to cultivated viruses was surprising. PSV was isolated from the 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 those of the springs in this study (71), and life within it includes a highly diverse population of *Archaea* and *Bacteria* (7, 39), most of which have not been detected in the

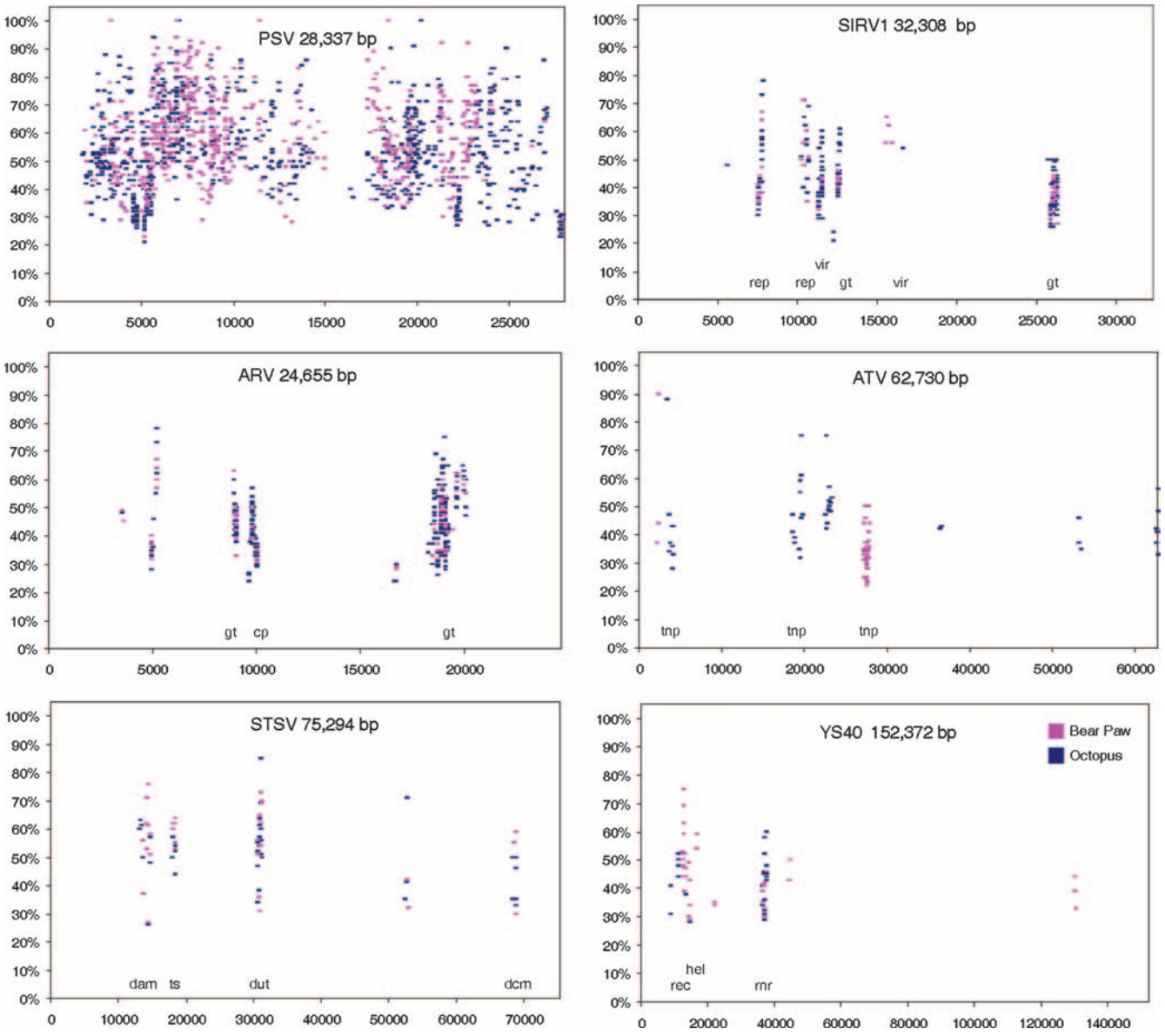

FIG. 6. Alignment of Octopus and Bear Paw viral metagenomic library contigs with six cultured virus genomes. Contigs assembled at >95% identity from the viral metagenomic libraries were compared by tBLASTx to the genomes of PSV, SIRV1, ARV, ATV, STSV, and YS40. Each bar represents the alignment of a unique metagenomic sequence to the indicated location on the cultivated viral genome, shown on the horizontal axis. Percent coding sequence identities are shown on the vertical axis. The threshold for inclusion of a contig is an E value of $<10^{-3}$. The red bars indicate Bear Paw alignments; the 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; D. Prangishvili, personal communication).

Octopus hot spring (10, 64) or elsewhere. In contrast, *Thermoproteus tenax spherical virus*, which is quite similar to PSV in terms of sequence, morphology, and habitat (1), had very limited similarity to the YNP viral metagenomic sequences (not shown). The 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 more remarkable because the microbial populations of acidic and neutral hot springs are quite distinct (66). The one other virus cultivated from YNP, *Sulfolobus spindle-shaped virus* Rabbit Hills (85), had no significant tBLASTx similarity to any of the metagenomic samples.

Similarities between the two hot-spring 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-

TABLE 6. Nucleotide and coding similarities between the viral populations of Octopus and Bear Paw hot springs

Parameter	Value	
	tBLASTx	BLASTn
Frequency (%) (no.) of Octopus contigs with similarity to Bear Paw contigs	43 (5,843)	21 (2,876)
Frequency (%) (no.) of Bear Paw contigs with similarity to Octopus contigs	26 (1,593)	21 (1,339)
Avg length of similarity (nucelotides)	298	175
Avg identity (%)	74	87
Avg expect value	1.38E−05	3.00E−05

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%) had amino acid coding similarity. By BLASTn, 2,876 (21%) and 1,339 (21%) of the respective contigs had nucleotide similarity. The average identities were 74% and 87%, and the expect values were 1.38E−05 and 3.00E−05, although the average lengths of sequence alignment (298 and 175 bp) were 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 degrees of identity at the 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. 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 the hot springs differ by 19°C (74°C versus 93°C).

Conservation and distribution of viruses in thermal environments. Taken together, the tBLASTx alignments suggest that (i) viral populations in the water columns are largely independent of microbial populations in the pools and (ii) viral genomes, particularly certain genes, are more conserved both regionally and globally than might have been predicted. 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, 72). 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 of 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 are 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 kilometers 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 contributions 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, collected within seconds of their emergence, have viral abundances similar to or somewhat higher than those measured in this report (16), suggesting that subsurface proliferation is at least a significant contributor to viral populations at the surface. Subsurface proliferation of viruses would also explain the apparent disconnect between the planktonic viral populations in the pool and the reported sedentary microbial populations described above. An implication of subsurface proliferation of viruses is that 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 for life in the subterranean aquifers could significantly exceed the temperatures measured at the surface.

ACKNOWLEDGMENTS

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 Mitchell, Ann Rodman, Carrie Guiles, Christie Hendrix, and John Varley for help with obtaining permits 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. The 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 T.S., by grant DOE DE-FG02-02ER83484 to D.M., and by 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 generation 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, 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 Bioinform. **6:**41.
5. **Angly, F. E., 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:**e368.

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. Rev. **70:**876–887.
9. **Bench, S. R., T. E. Hanson, K. E. Williamson, D. Ghosh, M. Radosovich, K. Wang, and K. E. Wommack.** 2007. Metagenomic characterization of Chesapeake Bay virioplankton. Appl. Environ. Microbiol. **73:**7629–7641.
10. **Blank, C. E., 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:**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:**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. Bacteriol. **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. Springer-Verlag, New York, NY.
18. **Brown, J. R., and W. F. Doolittle.** 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. **61:**456–502.
19. **Büchen-Osmond, C.** 2003. Taxonomy and classification of viruses, p. 1217–1226. *In* P. R. Murray, E. J. Baron, J. H. Jorgensen, M. A. Pfaller, and R. H. Yolken (ed.), Manual of clinical microbiology, 8th ed., vol. 2. ASM Press, Washington, DC.
20. **Canchaya, C., G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann, and H. Brussow.** 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. **4:**69.
24. **Daubin, V., and H. Ochman.** 2004. Start-up entities in the origin of new genes. Curr. Opin. Genet. Dev. **14:**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:**763–773.
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 evolutionary 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:**793–803.
31. **Fournier, R. O.** 2005. Geochemistry and dynamics of the Yellowstone National Park hydrothermal system, p. 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, p. 55–75. *In* J. Kieleczawa (ed.), DNA sequencing: optimizing the process and analysis. Jones and Bartlett, Sudbury, MA.
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:**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:**233–242.
37. **Hatfull, G. F., M. L. Pedulla, D. Jacobs-Sera, P. M. Cichon, A. Foley, M. E. Ford, R. M. Gonda, J. M. Houtz, A. J. Hryckowian, V. A. Kelchner, et al.** 2006. Exploring the mycobacteriophage metaproteome: phage genomics as an educational platform. PLoS Genet. **2:**e2.
38. **Hjörleifsdottir, S. H., G. O. Hreggvidsson, O. H. Fridjonsson, A. Aevarsson, and J. K. Kristjansson.** December 2002. U.S. patent 6,492,161.
39. **Hugenholtz, P., C. Pitulle, K. L. Hershberger, and N. R. Pace.** 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:**655–663.
42. **Kessler, A., A. B. Brinkman, J. van der Oost, and D. Prangishvili.** 2004. Transcription of the rod-shaped viruses SIRV1 and SIRV2 of the hyperthermophilic archaeon sulfolobus. J. Bacteriol. **186:**7745–7753.
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. **Lucchini, S., F. Desiere, and H. Brussow.** 1998. Comparative genomics of *Streptococcus thermophilus* phage species supports a modular evolution theory. Virology **246:**63–73.
46. **Lukashin, A., and M. Borodovsky.** 1998. GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res. **26:**1107–1115.
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:**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. U.S. Geological Survey, Boulder, CO.
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, 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 Microbiol. Lett. **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:**226–234.
56. **Peng, X., A. Kessler, H. Phan, R. A. Garrett, and D. Prangishvili.** 2004. Multiple variants of the archaeal DNA rudivirus SIRV1 in a single host and a novel mechanism of genomic variation. Mol. Microbiol. **54:**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:**52–67.
58. **Prangishvili, D., and R. A. Garrett.** 2004. Exceptionally diverse morphotypes

and genomes of crenarchaeal hyperthermophilic viruses. Biochem. Soc. Trans. **32:**204–208.
59. **Prangishvili, D., K. Stedman, and W. Zillig.** 2001. Viruses of the extremely thermophilic archaeon *Sulfolobus*. Trends Microbiol. **9:**39–43.
60. **Prangishvili, D.** 2003. Evolutionary insights from studies on viruses of hyperthermophilic archaea. Res. Microbiol. **154:**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, p. 345–421. *In* J. T Staley and A.-L. Reysenbach (ed.), Biodiversity of microbial life. Wiley Liss, New York, NY.
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:** 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:**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. Bacteriol. **185:**6434–6447.
71. **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.
72. **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:**480–486.
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:**561–566.
74. **Snyder, J. C., K. Stedman, G. Rice, B. Wiedenheft, J. Spuhler, and M. J. Young.** 2003. Viruses of hyperthermophilic archaea. Res. Microbiol. **154:** 474–482.
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:**801–812.
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:**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:**1353–1370.
83. **Weinbauer, M. G., and F. Rassoulzadegan.** 2004. Are viruses driving microbial diversification and diversity? Environ. Microbiol. **6:**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:** 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. **Yu, M. X., M. R. Slater, and H. W. Ackermann.** 2006. Isolation and characterization of *Thermus* bacteriophages. Arch. Virol. **151:**663–679.