JOURNAL OF VIROLOGY, Feb. 2004, p. 1954–1961 Vol. 78, No. 4
0022-538X/04/$08.00+0 DOI: 10.1128/JVI.78.4.1954–1961.2004

Comparative Genomic Analysis of Hyperthermophilic Archaeal *Fuselloviridae* Viruses

Blake Wiedenheft,[1,2] Kenneth Stedman,[1,3] Francisco Roberto,[4] Deborah Willits,[1] Anne-Kathrin Gleske,[1] Luisa Zoeller,[3] Jamie Snyder,[1,2] Trevor Douglas,[1,5] and Mark Young[1,2]*

Thermal Biology Institute,[1] Department of Microbiology,[2] and Department of Chemistry,[5] Montana State University, Bozeman, Montana 59717; DOE-INEEL, Idaho Falls, Idaho 83415[4]; and Department of Biology, Portland State University, Portland, Oregon 97201[3]

Received 15 August 2003/Accepted 20 October 2003

The complete genome sequences of two *Sulfolobus* spindle-shaped viruses (SSVs) from acidic hot springs in Kamchatka (Russia) and Yellowstone National Park (United States) have been determined. These nonlytic temperate viruses were isolated from hyperthermophilic *Sulfolobus* hosts, and both viruses share the spindle-shaped morphology characteristic of the *Fuselloviridae* family. These two genomes, in combination with the previously determined SSV1 genome from Japan and the SSV2 genome from Iceland, have allowed us to carry out a phylogenetic comparison of these geographically distributed hyperthermal viruses. Each virus contains a circular double-stranded DNA genome of ~15 kbp with approximately 34 open reading frames (ORFs). These *Fusellovirus* ORFs show little or no similarity to genes in the public databases. In contrast, 18 ORFs are common to all four isolates and may represent the minimal gene set defining this viral group. In general, ORFs on one half of the genome are colinear and highly conserved, while ORFs on the other half are not. One shared ORF among all four genomes is an integrase of the tyrosine recombinase family. All four viral genomes integrate into their host tRNA genes. The specific tRNA gene used for integration varies, and one genome integrates into multiple loci. Several unique ORFs are found in the genome of each isolate.

Comparative genomics is a useful tool for understanding new viral families. One such family, the *Fuselloviridae*, has recently been created to accommodate the approximately 60- by 90-nm spindle-shaped viruses found exclusively in the archaeal domain (International Committee on Taxonomy of Viruses, http://www.ncbi.nlm.nih.gov/ICTV/). This viral family presently consists of a single virus, *Sulfolobus* spindle-shaped virus 1 (SSV1), with three others considered tentative members in this genus: SSV2, SSV3, and the satellite virus pSSVx (for plasmid SSV x) (42). These viruses have circular double-stranded DNA genomes, share a common morphology and are temperate in *Sulfolobus* species that commonly inhabit high-temperature (>70°C), acidic (pH of <4.0) environments (43). *Haloarcula hispanica* 1 (5), *Sulfolobus neozealandicus* droplet-shaped virus (4), and a virus-like-particle isolated from *Methanococcus voltae* A3 (40) are morphologically similar to the *Fuselloviridae*. However, their genome topology, genomic structures, and host ranges vary dramatically, making their formal classification unclear.

SSV1 is the type virus of the *Fuselloviridae* family and the first high-temperature virus to be characterized in detail. SSV1 was originally isolated from *Sulfolobus shibatae* cultured from a sulfurous hot spring in Beppu, Japan (10, 17, 22, 41). The virus can also infect virus-free strains of *Sulfolobus solfataricus* originally isolated from a solfataric field near Naples, Italy (31, 44). In both hosts, virus production is UV inducible (17, 31) and the genome is stably maintained in three different forms. The packaged viral genome is positively supercoiled, while the episomal form of the viral genome has been isolated from *Sulfolobus* as positively supercoiled, negatively supercoiled, or relaxed double-stranded DNA (20). A provirus is also found integrated into a host tRNA gene (25, 27, 41). A 7.4-kbp segment inserted into an *S. solfataricus* arginyl tRNA gene shares extensive sequence similarity with a portion of the SSV1 genome and is likely a remnant of viral integration (25). The low G+C content (39.7%) of the viral genome is similar to that of its host. Sequence analysis of the encapsidated 15,465-bp genome (accession number NC_001338) revealed 34 open reading frames (ORFs) (Fig. 1 and Table 1) (22). The predicted ORFs encode protein products that range from 6 to 86 kDa and are tightly arranged on the viral genome. Nine transcripts cover all 34 SSV1 ORFs (28). This suggests that the viral genes are translated via a polycistronic strategy. Only 4 of the 34 ORFs have been assigned a function. The ORFs encoding three structural proteins, VP1, VP2, and VP3, were assigned by sequencing proteins from purified virus particles (26), while the fourth ORF was recognized to encode a type I tyrosine recombinase family integrase by sequence similarity to other known recombinases and by direct biochemical analysis (18, 19, 22, 32). The remaining 30 ORFs show little or no significant sequence similarity to genes in the public databases. SSV1-based shuttle vectors that can replicate in both *S. solfataricus* and *Escherichia coli* have been described (8, 14, 34). These vectors have the potential to greatly expand our ability to express genes in *S. solfataricus.*

The complete genome sequence of SSV2 has recently been determined (accession number AY370762) (Fig. 1) (35). This

* Corresponding author. Mailing address: Montana State University, Dept. of Plant Sciences and Plant Pathology, P.O. Box 173150, Bozeman, MT 59717. Phone: (406) 994-5158. Fax: (406) 994-7600. E-mail: myoung@montana.edu.

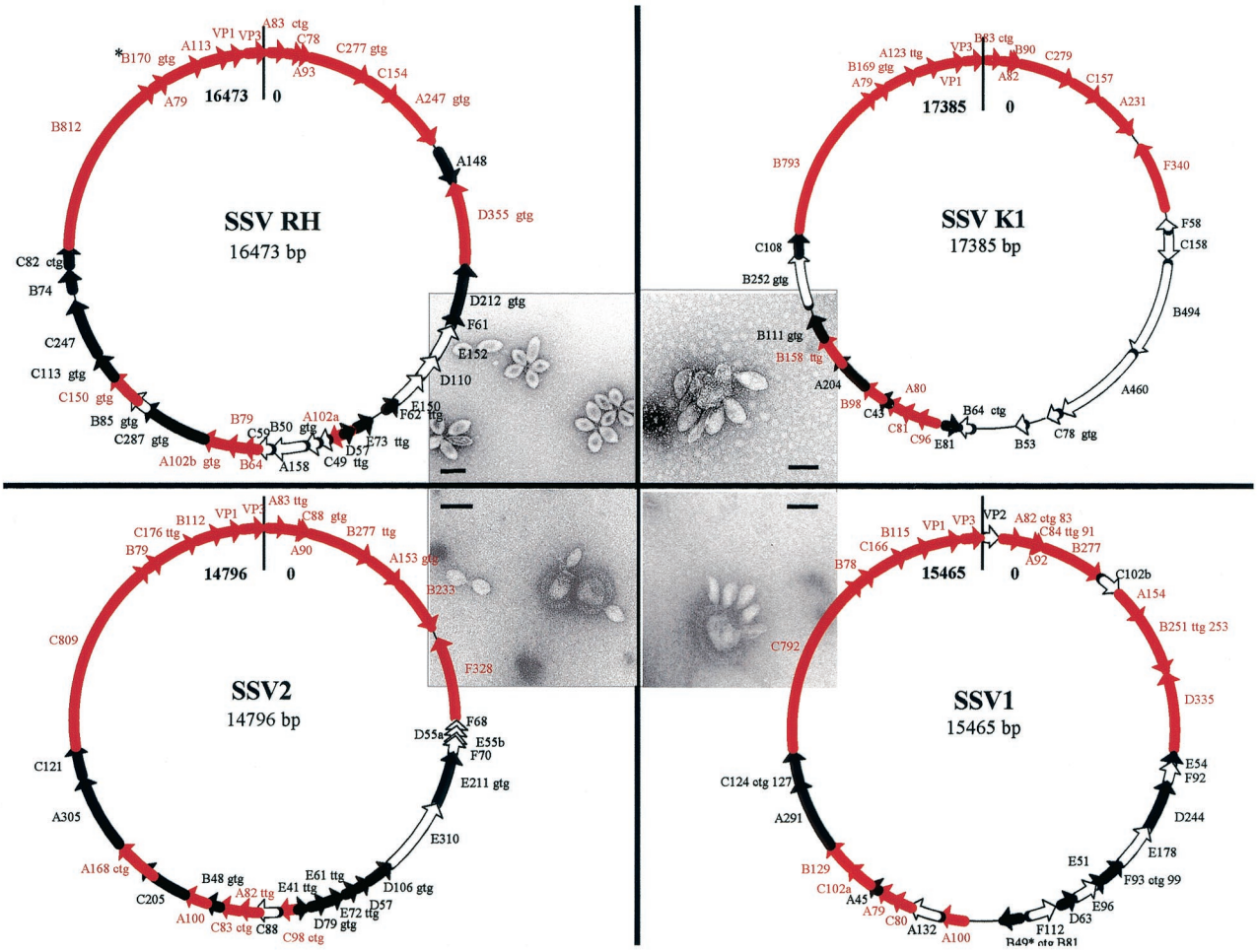

FIG. 1. Genome maps of the four isolates. Conserved ORFs shared by all genomes are shown as red arrows. ORFs shared between two or three of the SSV genomes are shown as solid black arrows, and ORFs unique to each isolate are shown as open arrows. TEM images of each virus are positioned next to maps of their respective genome (bars, 100 nm) (3, 34). The alternative initiation codons (asterisks) are indicated directly following the name of each ORF in which they were identified.

virus was isolated from a solfataric hot spring in Reykjanes, Iceland, and has been shown to infect *S. solfataricus* (3). The SSV2 virus is morphologically indistinguishable from SSV1, and the two viral genomes are similarly arranged. Overall, the SSV2 and SSV1 genomes share 26 ORFs, while 9 ORFs are unique to SSV2 and 8 are unique to SSV1. SSV2, unlike SSV1, does not encode a VP2 structural protein. The genomic comparison of SSV2 with SSV1 indicates that these are two distinct *Fuselloviridae* viruses (35). The original culture containing SSV2 also produced a subset of smaller satellite particles (60 by 40 nm) termed pSSVx (accession number AJ243537). The genetic material packaged by the satellite particle is a hybrid of pRN family plasmids and two ORFs that share sequence similarity to ORFs in SSV1 and SSV2 (3, 15, 16).

Full-length copies of both SSV1 and SSV2 genomes are found integrated into tRNA genes of their host (27, 41). The virally encoded integrase identified in the SSV1 (ORF D335) genome has been shown to function as a site-specific endonuclease and ligase (18, 19, 32). The SSV1 integrase is also capable of excisive recombination in vitro, but the mechanism of excision in vivo is not presently understood (18, 22). The viral integrase, like all 130 members of the tyrosine recombinase family, contains the conserved RHRY tetrad (2, 21). The SSV1 genome is found integrated into an arginine tRNA gene (27), while SSV2 is thought to integrate into a glycyl tRNA gene (35). The process of viral integration is modeled to involve base pairing of the *att*A sequence found within the target tRNA gene with the *att*P site found within the viral integrase gene. In the case of SSV1, the *att*A site is a 44-bp region that is centered around the anticodon loop of an arginyl tRNA gene (27). The *att*P site is an identical sequence located within the 5′ half of the viral integrase gene. The SSV1 viral genome integrates in a way that complements the tRNA gene and is presumed to maintain tRNA function (27). The recently described SSV2 genome also contains a tyrosine recombinase-like integrase gene (ORF F328) (35). Forty-nine nucleotides in this gene are identical to sequences found in a glycyl tRNA gene in the *S. solfataricus* P2 genome (33). These regions likely represent the *att*A and *att*P sites for SSV2. Currently, fuselloviruses offer the only viral model for examining site-specific recombination in archaea. The biological role of SSV integration and the in vivo mechanisms for SSV excision remain unclear.

TABLE 1. Identified ORFs in SSV isolates

ORFs[a] in:			
SSV RH (n = 37)	SSV K1 (n = 31)	SSV2 (n = 35)	SSV1 (n = 34)[b]
			VP2 (B74)
A83 (TTG)	**B83 (CTG)**	**A83 (TTG)**	**A82 (CTG), 83**
C78	**A82**	**C88**	**C84 (TTG), 91**
A93	**B90**	**A90**	**A92**
C277 (GTG)	**C279**	**B277 (TTG)**	**B277**
			C102b
C154	**C157**	**A153 (GTG)**	**A154**
A247 (GTG)	**A231**	**B233**	**B251 (TTG), 253**
A148			
D355 (GTG)	**F340**	**F328**	**D335**
	F58		
	C158		
	B494		
	B460		
	C78 (GTG)		
	B53		
	B64 (CTG)		
			E54
			F92
		F68	
		D55a	
		E55b	
		F70	
D212 (GTG)		E211 (GTG)	D244
E152			
D110			
E150			
			E178
		E310	
	E81	D106 (GTG)	F93 (CTG), 99
			E96
F61		D57	D63
			F112
F62 (TTG)		E61 (TTG)	
		E72 (TTG)	B49 (CTG), 81
E73 (TTG)		D79 (GTG)	E51
A102a	**C96**	**C98 (CTG)**	**A100**
D57		E41 (TTG)	
C49 (TTG)			
B50 (GTG)			
A158			
C59			
		C88	
			A132
B64	**C81**	**A82 (TTG)**	**C80**
B79	**A80**	**C83**	**A79**
	C43	B48 (GTG)	A45
A102b (GTG)	**B98**	**A100**	**C102a**
C287 (GTG)	A204	C205	
B85 (GTG)			
C150 (GTG)	**B158 (TTG)**	**A168 (CTG)**	**B129**
C113 (GTG)	B111 (GTG)		
	B252 (GTG)		
C247		A305	A291
B74			
C82 (CTG)			
	C108	C121	C124 (CTG), 127
B812	**B793**	**C809**	**C792**
A79	**A79**	**B79**	**B78**
B170 (GTG)	**B169 (GTG)**	**C176 (TTG)**	**C166**
A113	**A123 (TTG)**	**B112**	**B115**
VP1 (A89)	**VP1 (B137)**	**VP1 (C88)**	**VP1 (C144)**
VP3 (C96)	**VP3 (A93)**	**VP3 (A92)**	**VP3 (A92)**

[a] Homologous ORFs are in the same row, and ORFs that are common to all four SSV genomes are in boldface. Alternative initiation codons identified are in parentheses.

[b] Alternative initiation codons were not considered in the original SSV1 annotation. The originally annotated names are included here so as not to confuse them with the new names (numbers) that we identified by using the alternative initiation codon indicated.

SSVs and their *Sulfolobus* hosts are emerging as a model system for examining archaea and life at high temperatures. We are interested in the evolution of SSVs, the function of SSV-encoded gene products, and viral adaptations required for replication in high-temperature environments. Here we present the genomes of two additional SSV-like viruses, one from Kamchatka, Russia (SSV K1), and the other from Yellowstone National Park, United States (SSV RH). These genomes in combination with the previously determined SSV1 and SSV2 genomes provide us with four geographically distinct isolates that we have used in a comparative genomic analysis. In addition, we have determined the sites of integration of the two new genomes into *S. solfataricus*. This work is the first such comparative analysis within the archaeal domain.

MATERIALS AND METHODS

Environmental sampling. Liquid samples were collected from an acidic hot spring (pH 4.0, 75°C) at 54°26.357′N, 160°8.573′E in the Valley of the Geysers, Kamchatka, Russia. The pH of the Kamchatka samples was adjusted to 5.0, and the samples were stored in anaerobic vials for transport as described previously (43). Liquid samples collected from acid chloride hot springs (pH 3.2, 81°C) in the Norris Geyser Basin of Yellowstone National Park, United States (44°43.653′N, 110°42.862′W), were transported aerobically and placed in enrichment culture within 6 h as previously described (29).

Enrichment cultures of environmental samples. Enrichment cultures were established by inoculating 20 ml of minimal medium supplemented with 0.1% tryptone and adjusted to pH 3.2 (43) with 1 ml of environmental sample. Liquid cultures were grown aerobically in long-neck Erlenmeyer flasks placed in shaking oil bath incubators at 80°C for 5 to 7 days. Turbid cultures were streaked on 0.6% Gelrite gellan gum plates supplemented with 0.2% tryptone (43). Single-colony clones were isolated and screened for plaque formation as described previously (34, 43). Single-colony isolates were used to establish 25-ml cultures. Seven days after inoculation, culture supernatants were visually screened by transmission electron microscopy (TEM) (Leo 912 AB) for the presence of virus-like particles as previously described (29). Cultures found to be producing large quantities of virus-like particles were scaled up to 250-ml cultures. Total DNA was extracted from 1.5-ml aliquots (34). This DNA was used as a template for PCR-based amplification of the 16S rRNA gene, which was subsequently cloned and sequenced as previously described (29).

Virus purification and viral nucleic acid isolation. Cells were isolated from cultures by low-speed centrifugation, and their extrachromosomal DNA was isolated by previously described methods (34, 41). Virus was precipitated from culture supernatants by addition of polyethylene glycol 8000 (10% [wt/vol] final concentration) and stirred at 4°C overnight. The precipitated virus was collected by low-speed centrifugation and resuspended in a minimal volume of NNM buffer (20 mM $NaPO_4$ [pH 5.5], 100 mM NaCl, 1 mM $MgCl_2$). After low-speed centrifugation to remove material that did not resuspend, the supernatant was mixed with CsCl to a final concentration of 39% (wt/vol) and spun at 69,000 × *g* for 24 h in an SW 41 rotor (Beckman, Fullerton, Calif.). The dominant band was removed and dialyzed against NNM buffer overnight at 4°C. Virus was analyzed by TEM and UV-visible spectroscopy. Nucleic acid was isolated from purified virus by the sodium dodecyl sulfate-proteinase K method (30).

Construction of viral DNA library and sequencing. DNA was mechanically sheared by nebulization as outlined by the manufacturer (Shotgun cloning kit; Invitrogen, San Diego, Calif.). Sheared DNA was ligated into pCR 4Blunt-TOPO (Invitrogen) and transformed into *E. coli* XL-2 MRF′ (Stratagene, San Diego, Calif.). Plasmid DNAs from colonies containing inserts of greater than 500 bp were prepared for sequencing according to the instructions of the manufacturer (Eppendorf, Perfect Preps, Westbury, N.Y.). Purified DNA was sequenced by using universal M13 primers according to Big Dye III termination sequencing protocols on an ABI 3700 automated DNA sequencer (Applied Biosystems, Foster City, Calif.). After sequence assemblies (see below), any remaining gaps were PCR amplified with primers that flanked each gap. These PCR products were subsequently cloned and sequenced to provide at least threefold sequence coverage of the genome, except for one small region of about 100 nucleotides (nt) with only twofold coverage. In this instance, both strands are sequenced and of a high quality.

Sequence analysis. Vector stripping and contig assemblies were accomplished with Sequencher version 4.1 (Gene Codes Corp., Ann Arbor, Mich.). ORFs were

initially identified by using ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi). ORFs were confirmed and others were identified by manually scanning for TTG, GTG, or ATG start codons between stop codons in all six frames. All possible ORFs were subjected to BLAST analysis against the nonredundant (GenBank) database (http://www.ncbi.nlm.nih.gov/BLAST/) (1). ORFs predicted to encode products of fewer than 50 amino acids and having no significant match to the database were discarded. Small ORFs (<50 amino acids) that have sequence similarity to ORFs in other SSV genomes were subjected to further analysis. Additional ORFs (<100 amino acids) were eliminated if they had no homologue in the database or in the other SSV genomes and if they overlapped a larger ORF by more than 50%. Large ORFs (>100 amino acids) having no SSV homologue were allowed to overlap by up to 30 amino acids before the smaller ORF was eliminated. The SSV RH ORFs identified by this method were compared to those ORFs predicted by Glimmer 2.0 (9) and RBS finder (36). Numbering of the SSV RH and SSV K1 genomes was standardized by using the first nucleotide following the stop codon of VP3 as nucleotide one and the last nucleotide in the stop codon of VP3 as the final nucleotide of the genome. All ORFs in the frame starting with the +1 nucleotide are designated A, those starting with +2 are designated B, those starting with +3 are designated C, and those on the opposite strand are designated D, E, and F, respectively. ORFs of the same length and in the same frame are distinguished by a lowercase letter (e.g., A102a). In Table 1 all identified ORFs are listed in the column below their respective genomes, and homologous ORFs are found in rows. Tandem repeats were identified by using Tandem Repeats Finder (http://c3.biomath.mssm.edu/trf.html) (6).

SSV genomic comparisons. TFASTX was used to compare each predicted gene product to the six-frame translations of all four SSV genomes (Biology Workbench [http://workbench.sdsc.edu/]). Predicted gene products sharing sequence similarity were considered to be possible protein homologues and were subjected to further alignments by using CLUSTALW (38). If amino acid identities between sequences were less then 25%, then the lengths of the two proteins, their coding directions, and their locations on the genome were considered. Colinear proteins (found in at least two of the four genomes) of similar length were considered functionally equivalent genes.

Phylogenetic comparisons of common ORFs. Amino acid sequences of each of the eighteen common SSV ORFs were aligned by using CLUSTALX (13). Maximum-parsimony and neighbor-joining analyses were conducted with test version 4.0b10 of PAUP* (37). Bootstrap analysis with resampling was performed on 10,000 replicates in each analysis. Maximum-likelihood analysis was performed on a concatemer consisting of all of the nucleotides in the set of 18 conserved ORFs from each genome. MrBayes analysis was run by sampling every 10,000 generations until the chain reached apparent stationarity (11, 12).

Analysis of sites of SSV integration into the *S. solfataricus* genome. Potential viral genome integration sites were predicted in a two-step process. First, the integrase ORFs containing the presumed *att*P sequences were individually aligned to a database containing all of the annotated tRNA genes in the *S. solfataricus* P2 genome (33) by using FASTA alignments (http://workbench.sdsc.edu/). From these alignments, putative *att*P sites were identified and aligned to the entire *S. solfataricus* P2 genome by using BLASTN (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/geblast?gi = 180) (1). Sets of PCR primers were designed to flank all regions where the putative viral *att*P site aligned to the *S. solfataricus* P2 sequence with an *e* value of less than $2.0e^{-09}$ (Table 2). A PCR-based integration assay was designed such that one PCR primer corresponds to the viral sequence near the putative site of integration and a second primer flanks the tRNA gene of interest. Sites that were not annotated as tRNA genes were tested by using the same strategy. Fifty-milliliter cultures of *S. solfataricus* P2 were grown to mid-log phase and then infected with filtered (0.2-μm-pore-size filter) supernatants of either SSV RH- or SSV K1-containing cultures. Infected cultures were then grown to stationary phase, at which point the cells were harvested by low-speed centrifugation (6,000 × *g* for 10 min) and total DNA was prepared as previously described (34). This DNA was used as the template in PCR-based integration assays. PCRs were cycled 35 times, and the annealing temperatures were generally 5°C below the lowest predicted melting temperature in the primer set. The resulting PCR products were sequenced as described above.

Structural modeling of tRNAs. tRNAscan (http://www.genetics.wustl.edu/eddy/tRNAscan-SE/) was used to model the structure of *S. solfataricus* tRNAs before and after integration.

TABLE 2. Primers used in PCR-based assay of integration of SSVRH and SSV K1 into P2

Primer	Sequence
SSV RH	
Right side	
SSVRH (virus)	GGATTCGTGAGGTTAAGGGG
P2 L1 (host)	GCTTAGAGATGGAACTGCACCCC
P2 L2 (host)	CGCATTCATCCATGTAACC
P2 L3 (host)	CCATCTGGAACGTTGTTTCC
P2 L4 (host)	GGCGTTAAAGAGGTTATGG
P2 L5 (host)	GAGCTTCTTAACTCCGTTCCTTCC
P2 R1 (host)	GCAACCGGAAAACCTTCTCC
Left side	
RH (virus)	CACGCGTGATTTCCATGTCC
P2 L5 (host)	CGCTTTCAGCTATTAGCGGGG
SSV K1	
Right side	
SSV K1 (virus)	CTCAGAGGGCGGATCTCTG
P2 D (host)	GGGAAACCCCGAGGTCCCTGG
P2 E1 (host)	CTGAAGTACAAATGTCAGCG
P2 E2 (host)	GGTTATTGTGAGGGATGTAGAGG
Left side	
SSV K1 a (virus)	GCCTAGTTTCTATGTCGG
SSV K1 b (virus)	GCCGTCTTCTTTCAATTTCTTTAC
P2 D (host)	CATTCTAACTCCTTCCTCGC
P2 E1 (host)	CCCCACGTAAATTACATTC
P2 E2 (host)	CCTACCTATACCTAACTTCTGTGC
P2 1632500 (host)	GGAAAGGTGGATAGCTAAATTGCGC

RESULTS AND DISCUSSION

We have isolated two new SSV-like viruses and determined their genome sequences. SSV RH (for Ragged Hills) was isolated from an acid chloride (pH 3.2) hot spring (81°C) in the Ragged Hills region of the Norris Geyser basin in Yellowstone National Park (United States). SSV K1 (for Kamchatka 1) was isolated from a hot (75°C) acidic (pH 4.0) pool in the Geyser Valley region of the Uzhno-Kamchatsky National Park on the Kamchatka peninsula (Russia). Small-subunit ribosomal DNA sequence analysis indicated that the hosts of both viral isolates are closely related to *S. solfataricus*. In addition to their natural hosts, these viruses can also infect virus-free strains of *S. solfataricus* P2 (from Pisciarelli, Italy), like the previously characterized SSV1 and SSV2 viruses (3, 31, 35).

All of these viruses share a unique spindle-shaped morphology that has been seen only in the archaeal domain. These virus particles are all about 60 by 90 nm with sticky tail fibers extending from one end. These tail fibers are presumed to be involved in viral attachment to the host and association with membrane vesicles and are likely the cause of virus clustering into rosette formations as seen in culture (Fig. 1). Although the spindle-shaped morphology is predominant, the virus structure appears to be malleable and is also able to form elongated or cigar-shaped morphologies (3, 17, 31, 43).

SSV RH and SSV K1 virions contain double-stranded circular DNA genomes of approximately 15 kbp. The entire genome sequences of both viruses were determined by random shotgun sequencing. The genome of SSV RH is 16,473 bp (accession number AY388628) and that of SSV K1 is 17,384 bp (accession number AY423772), both of which are larger than the SSV1 (15,465 bp) and SSV2 (14,796 bp) genomes (22, 35). All four viral genomes have a G+C content of ~38%, like their *S. solfataricus* host (Fig. 1 and Table 1).

Genome analysis of these two new SSV-like viruses in combination with previously sequenced isolates indicates a clear relationship among all four viral isolates (Fig. 1 and Table 1). Thirty-seven ORFs have been identified in SSV RH, while the larger SSV K1 genome contains only 31 ORFs (Fig. 1 and Table 1). In comparison, 34 and 35 ORFs were identified in the SSV1 and SSV2 genomes, respectively. This is one more ORF than was previously reported for SSV2 (35). The additional ORF in SSV2, ORF E41, was identified by having significant sequence identity (63% at the amino acid level) to ORF D57 in SSV RH. Eighteen of the SSV RH ORFs are shared by all four isolates (Fig. 1 and Table 1), five SSV RH ORFs are shared between two of the other three viral isolates, and three of the SSV RH ORFs are shared by only one of the other three SSV isolates. SSV RH contains 11 unique ORFs. Nine of these ORFs are found clustered in groups of two to four, while the two remaining ORFs are found as isolated ORFs. SSV K1 contains eight unique ORFs. Seven of these are found in a single cluster following the integrase gene. Three of the SSV K1 ORFs are shared with SSV1 and SSV2, while two SSV K1 ORFs are shared exclusively with SSV RH (Fig. 1 and Table 1).

A comparison of all four viral genomes reveals that each genome has a colinear organization (Fig. 1). As is typical of many viral genomes, the predicted ORFs are tightly arranged on the genome, with little noncoding sequence. There is little overlap between adjacent ORFs. SSV1 ORFs C84 and A92 overlap by 235 nucleotides and are the exception to this rule. Homologues of these two overlapping ORFs are found in all four viral genomes, indicating a functional role for this overlap or for the DNA sequence conserved in both ORFs (Fig. 1). ORFs F70, D55a, D55b, and F68 of SSV2 overlap by approximately 100 bp each, but these ORFs are not conserved in the other three SSV genomes. Like those of SSV1 and SSV2, the SSV RH and SSV K1 genomes encode potential products ranging from 5 to 90 kDa, with an average of ~16 kDa. Sequence alignments of the SSV ORFs show little or no sequence similarity to sequences in the public databases. However, among the four genomes, many ORFs share sequence similarity and are colinearly organized. In general, ORFs on one half of the genome are more highly conserved between the four isolates and are arranged in the same orientation, while genes on the other half of the genome are poorly conserved. The highest identity of a predicted ORF product shared by all isolates is 84% (SSV RH B170 homologues), and the lowest is 13% (SSV RH ORF A102a homologues). ORFs with low sequence identity were considered putative homologues only if they are of a similar size, in a similar location on the genome, and oriented in the same direction.

The asymmetric clustering of cysteine codons that has been observed in the SSV1 and SSV2 genomes is also obvious in the genomes of the two new SSV-like viruses. The consistent patterning of cysteine codons across all four SSV genomes strongly suggests that they share a common ancestor. It has been suggested that this ancestor was a fusion between two genomes with different histories, one that contained cysteine and another that did not (22, 35).

Three of the genes that lie on the conserved half of the genome have been assigned functions. Two are viral coat proteins (VP1 and VP3), and the third is a viral integrase (22, 26). VP1 and VP3 were identified by N-terminal sequencing of proteins isolated from purified SSV1 particles (26), and the integrase was identified by sequence similarity (2, 19, 22). The N-terminal sequence of purified VP1 was found to be identical to the C-terminal 73 amino acids predicted in SSV1 ORF C144. The absence of the N terminus indicates that VP1 is proteolytically processed (26). VP1 homologues have been identified in all four SSV genomes. The residues surrounding the putative proteolytic processing site are conserved in all four viral genomes, with the C-terminal ends of all four genes being highly conserved and the N-terminal ends being very divergent. The protease responsible for this processing has yet to be identified, but these alignments suggest a common processing mechanism of the VP1 proteins in all four SSV isolates.

VP2 is one of only three proteins identified in the mature SSV1 virion (26). VP2, like the other two structural proteins (VP1 and VP3), was identified by N-terminal sequencing of proteins isolated from purified virus particles. This protein is composed largely of basic residues and has been shown to be a DNA binding protein (26). The DNA binding activity and the presence of VP2 in assembled virus particles suggested that VP2 was an essential protein involved in packaging of the SSV1 viral genome. Surprisingly, there is no VP2 homologue in the other three SSV genomes or in the *S. solfataricus* (P2) genome (33). It remains unclear how these viruses package their genomes in the absence of a VP2 homologue. Interestingly, two ORFs (SSV1 A154 and B251 and their homologues in the other genomes) are also conserved in the satellite virus pSSVx (3). These ORFs probably encode gene products or contain sequences involved in packaging of the viral genomes; however, direct evidence for these possible functions is presently lacking. The predicted products encoded by the remaining 29 to 36 ORFs (depending on the isolate) share no significant similarity to proteins with known function.

A striking feature revealed by this comparison is a set of 18 ORFs common to all four SSV isolates. We propose that the products encoded by this set of 18 common ORFs may represent viral functions common to all fuselloviruses and clearly reflect a common evolutionary history, despite their geographic isolation. This set of common genes may also represent the minimal replicon of the *Fuselloviridae*. Maximum-parsimony and neighbor-joining trees generated for each of the 18 conserved ORFs do not result in a consistent branching pattern. However, a maximum-likelihood tree generated by using a concatemer of nucleotide sequences from all 18 ORFs in each SSV genome suggests that SSV1 and SSV K1 are more genetically similar, while SSV RH and SSV1 are the most genetically different.

This common set of 18 ORFs shared by all four SSV isolates is not contiguous. As discussed above, approximately half of the SSV genome is highly conserved among all four isolates (Fig. 1), while the other half is more divergent. ORFs not common to all four SSV genomes rarely interrupt ORFs in the conversed half of the genome. The other half of the genome is largely composed of ORFs that are common to only two or three of the SSV genomes (Fig. 1). ORFs in any one SSV genome that have no significant sequence similarity to ORFs in the other SSV isolates are also indicated in Fig. 1. These unique ORFs are likely the consequence of their individual evolutionary history, geographic isolation, requirements for

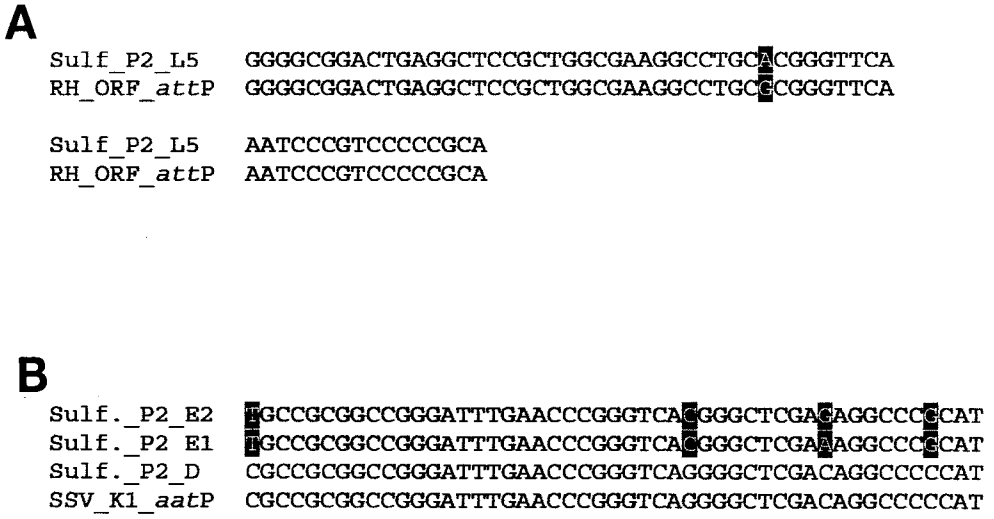

FIG. 2. Sequence alignments of the viral *att*P sites and host tRNA genes. (A) Alignment of the sequence around the SSV RH *att*P site and the fifth leucine (L5) tRNA gene of *S. solfataricus* P2. (B) Multiple alignment of the sequence around the SSV K1 *att*P site, the aspartic acid tRNA gene (D), the first glutamic acid tRNA gene (E1), and the second glutamic acid tRNA gene (E2). Sequence variations are indicated by shading.

replication in their specific hosts, or adaptations to unique features of their respective thermal environments.

Two separate imperfect direct repeats have been identified in the SSV RH genome, while no repetitive sequences were identified in the SSV K1 genome, using the same method. The first repeat (positions 6261 to 6337 in the SSV RH genome) has a consensus sequence of TTCTTCAGTTCTCAACAAC and occurs 3.9 times. The second repeat (positions 6243 to 6327) has a consensus sequence of TCTCACAACTTCTTCA GTTTCX and is repeated 3.7 times. Both of the sequences have a periodicity of 21 nt and are located between ORFs F61 and F62. These repeats are unique to SSV RH and are not related to repetitive sequences previously identified in the SSV1 or SSV2 genome (22, 35). The function or significance of these repetitive regions is presently not understood; they may be signal sequences involved in the recognition or regulation of viral replication.

We have identified tyrosine recombinase-like integrase genes in the genomes of the two new SSV isolates. Both SSV RH (ORF D355) and SSV K1 (ORF F340) viral integrase-like genes contain nucleotide sequences that are duplicated in tRNA genes in the *S. solfataricus* P2 genome (accession number NC_002754) (Fig. 2). These sequences represent potential *att*P and *att*A sites for directing integration of the viral genome. The integrase-like gene in SSV RH contains a contiguous stretch of 59 nucleotides that shares striking similarity to nucleotides in all five of the leucyl tRNA genes of *S. solfataricus* P2. Fifty-eight of these 59 nucleotides are exactly duplicated in the fifth leucyl tRNA (L5) gene, while this sequence is less conserved in the other four leucyl tRNA genes (Fig. 2A). The integrase-like gene in SSV K1 contains a contiguous stretch of 49 nt that is exactly duplicated in the only aspartic acid tRNA gene annotated in the *S. solfataricus* P2 genome. However, 45 of these 49 bases are duplicated in the *S. solfataricus* P2 glutamic acid tRNA genes E1 and E2 (Fig. 2B).

Based on these duplicated sequences, we assumed that the SSV RH and SSV K1 genomes should integrate into different tRNA genes of the *S. solfataricus* P2 genome. In an effort to characterize the occurrence and location of SSV RH and SSV K1 integration, we designed PCR primers that flank potential integration sites on the *S. solfataricus* P2 genome (Fig. 3 and Table 2). These primers were used together and in combination with viral primers that flank the putative *att*P sites. SSV integration is modeled in Fig. 3A, and the PCR-based approach to test for SSV integration is illustrated in Fig. 3B. PCR products produced by pairing a virus-specific primer with a host-specific primer were sequenced. SSV RH integrates site-specifically into the fifth leucyl tRNA gene of *S. solfataricus* P2 genome. Viral integration occurs at the exclusion of the other leucyl tRNA genes (data not shown). Viral integration occurs in such a way that the tRNA gene sequence is conserved except for a single T-to-C nucleotide change in the *S. solfataricus* P2 genome. This nucleotide is located at the 5′ end of the predicted T stem and may disrupt base pairing in the structure. SSV K1 integration appears to be more promiscuous. The SSV K1 genome integrates into three different tRNA genes: D, E1, and E2. All three of these tRNA genes contain putative *att*A sites that share considerable sequence similarity with the putative viral *att*P sequence (Fig. 2B). SSV K1 integration into the only aspartic acid tRNA gene in the genome results in an exact copy of the tRNA gene. However, SSV K1 genome integration into the glutamic acid tRNA genes E1 and E2 is not as precise. Integration at both locations eliminates the 5′ thymine from the tRNA genes and introduces a C-to-G mutation at a position corresponding to the first nucleotide of the anticodon stem of the tRNA. These mutations may not perturb gene function, but a second mutation (G to C in E2 and A to C in E1) that we frequently observed at the first position of the anticodon switches the anticodon from glutamic acid to aspartic acid. Variations in sequence at this location are a result of imprecise integration of the viral genome. Apparently the exact crossover point for integration of this viral genome in either the E1 or E2 tRNA gene is not always the same.

The viral and host genomes share sequences other than

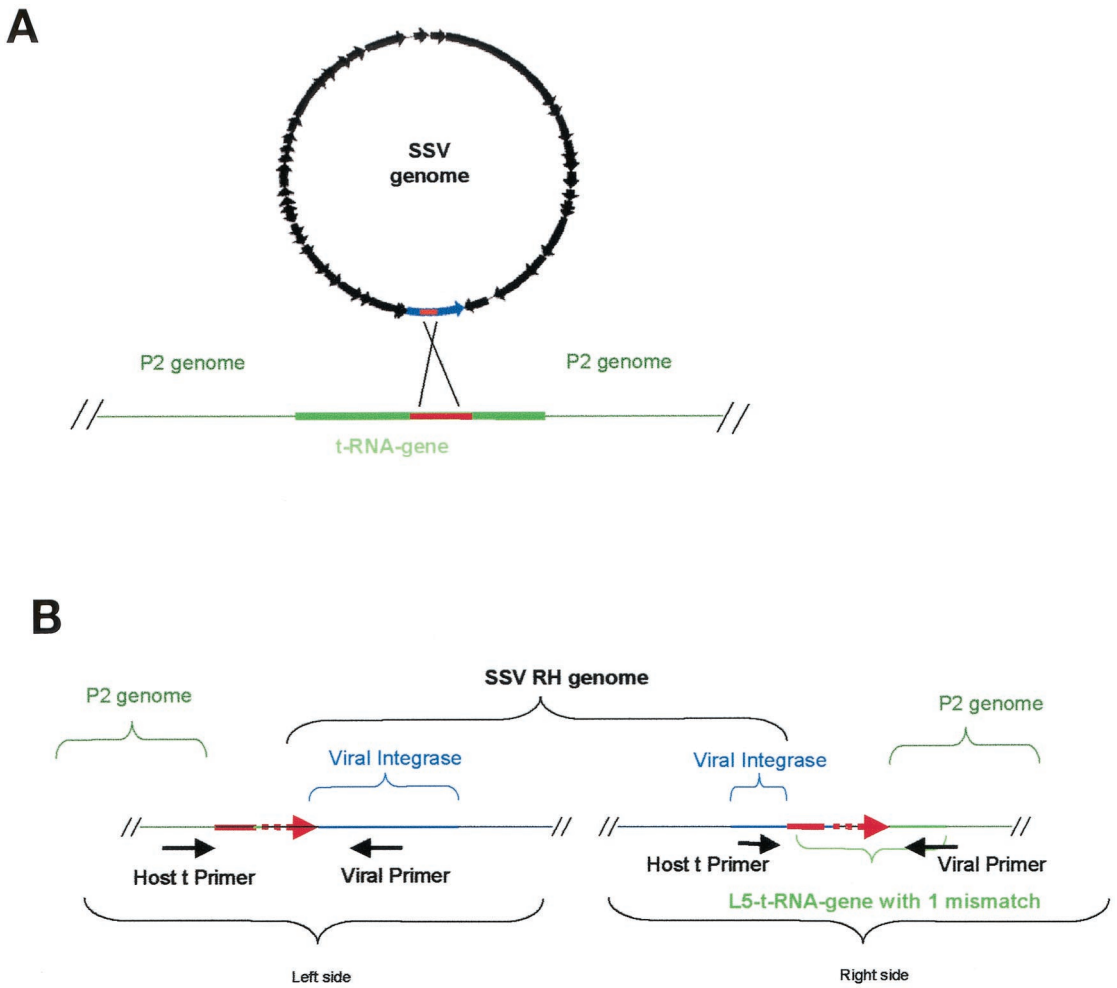

FIG. 3. Integration of SSV RH and SSV K1 into the *S. solfataricus* P2 genome. (A) General overview of SSV integration into a tRNA gene of *S. solfataricus* P2. The circular SSV genome is in black, the blue region represents the virus-encoded integrase, the green lines indicate host chromosome sequences, and the solid red lines indicate putative *att*P and *att*A sites in the virus and the host. (B) Schematic of SSV RH integrated into the host chromosome at the L5 tRNA gene. The general locations of viral integration and host primers that were used in the PCR-based integration assay are shown directly below the genomes. The dashed red lines indicate the region within which the recombination event occurs.

those identified as the putative *att*P and *att*A sites (24, 35). These similar sequences could potentially serve as alternative integration sites. We designed PCR primers to assay one of these alternative sites. In addition to integration at the tRNA sites, we have observed integration of the SSV K1 genome into a non-tRNA site located near position 1632500 in the *S. solfataricus* (P2) genome. This region shares 39 nt with the putative viral *att*P site. This is the first report of an SSV integrating at a non-tRNA location. This discovery was surprising, as bacteriophages that encode integrase proteins typically target tRNA genes, which has been considered an ancient process that has been conserved during the evolution of integrating viruses (7, 27).

The remarkable diversity observed in the genomes of these four clearly related SSV viruses is not presently understood. The fact that there is such diversity indicates that each of these isolates should be considered a different member of the *Fuselloviridae* and not a strain of the same virus. It is tempting to speculate that the observed diversity is a function of the geographic isolation of each virus, similar to that observed in their *Sulfolobus* hosts (39). However, first SSV viral diversity within an individual location must be established. The astonishing diversity observed in fuselloviruses is reminiscent of that in the tailed mycobacteriophages (23). SSV diversity may be a consequence of high turnover rates in viral populations, differences in viral hosts, and/or the chemical nature of the thermal features themselves. Regardless of the cause, the genome diversity that we have described should prove invaluable for guiding future experiments for understanding fuselloviruses and their environments.

ACKNOWLEDGMENTS

This work was supported by grants from the National Science Foundation (MCB01322156) and the National Aeronautics and Space Administration (NAG5-8807).

REFERENCES

1. **Altschul, S. F., T. L. Madden, A. A. Schaffer, 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.
2. **Argos, P., A. Landy, K. Abremski, J. B. Egan, E. Haggard-Ljungquist, R. H.**

Hoess, M. L. Kahn, B. Kalionis, S. V. Narayana, L. S. Pierson III, N. Sternberg, and J. M. Leong. 1986. The integrase family of site-specific recombinases: regional similarities and global diversity. EMBO J. **5:**433–440.
3. **Arnold, H. P., Q. She, H. Phan, K. Stedman, D. Prangishvili, I. Holz, J. K. Kristjansson, R. Garrett, and W. Zillig.** 1999. The genetic element pSSVx of the extremely thermophilic crenarchaeon Sulfolobus is a hybrid between a plasmid and a virus. Mol. Microbiol. **34:**217–226.
4. **Arnold, H. P., U. Ziese, and W. Zillig.** 2000. SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon *Sulfolobus.* Virology **272:** 409–416.
5. **Bath, C., and M. L. Dyall-Smith.** 1998. His1, an archaeal virus of the *Fuselloviridae* family that infects *Haloarcula hispanica.* J. Virol. **72:**9392–9395.
6. **Benson, G.** 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. **27:**573–580.
7. **Campbell, A. M.** 1992. Chromosomal insertion sites for phages and plasmids. J. Bacteriol. **174:**7495–7499.
8. **Cannio, R., P. Contursi, M. Rossi, and S. Bartolucci.** 1998. An autonomously replicating transforming vector for *Sulfolobus solfataricus.* J. Bacteriol. **180:** 3237–3240.
9. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
10. **Grogan, D., P. Palm, and W. Zillig.** 1990. Isolate B12, which harbours a virus-like element, represents a new species of the archaebacterial genus *Sulfolobus*, *Sulfolobus shibatae*, sp. nov. Arch. Microbiol. **154:**594–599.
11. **Huelsenbeck, J. P., and F. Ronquist.** 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17:**754–755.
12. **Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback.** 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science **294:**2310–2314.
13. **Jeanmougin, F., J. D. Thompson, M. Gouy, D. G. Higgins, and T. J. Gibson.** 1998. Multiple sequence alignment with Clustal X. Trends Biochem. Sci. **23:**403–405.
14. **Jonuscheit, M., E. Martusewitsch, K. M. Stedman, and C. Schleper.** 2003. A reporter gene system for the hyperthermophilic archaeon *Sulfolobus solfataricus* based on a selectable and integrative shuttle vector. Mol. Microbiol. **48:**1241–1252.
15. **Keeling, P. J., H. P. Klenk, R. K. Singh, O. Feeley, C. Schleper, W. Zillig, W. F. Doolittle, and C. W. Sensen.** 1996. Complete nucleotide sequence of the *Sulfolobus islandicus* multicopy plasmid pRN1. Plasmid **35:**141–144.
16. **Kletzin, A., A. Lieke, T. Urich, R. L. Charlebois, and C. W. Sensen.** 1999. Molecular analysis of pDL10 from *Acidianus ambivalens* reveals a family of related plasmids from extremely thermophilic and acidophilic archaea. Genetics **152:**1307–1314.
17. **Martin, A., S. Yeats, D. Janekovic, W.-D. Reiter, W. Aicher, and W. Zillig.** 1984. SAV1, a temperate U. V. inducible DNA virus like particle from the archaebacterium *Sulfolobus acidocaldarius* isolate B12. EMBO J. **3:**2165–2168.
18. **Muskhelishvili, G.** 1994. The archaeal SSV integrase promotes intermolecular excisive recombination in-vitro. Syst. Appl. Microbiol. **16:**605–608.
19. **Muskhelishvili, G., P. Palm, and W. Zillig.** 1993. SSV1-encoded site-specific recombination system in *Sulfolobus shibatae.* Mol. Gen. Genet. **237:**334–342.
20. **Nadal, M., G. Mirambeau, P. Forterre, W.-D. Reiter, and M. Duguet.** 1986. Positively supercoiled DNA in a virus-like particle of an archaebacterium. Nature **321:**256–258.
21. **Nunes-Duby, S. E., H. J. Kwon, R. S. Tirumalai, T. Ellenberger, and A. Landy.** 1998. Similarities and differences among 105 members of the Int family of site-specific recombinases. Nucleic Acids Res. **26:**391–406.
22. **Palm, P., C. Schleper, B. Grampp, S. Yeats, P. McWilliam, W.-D. Reiter, and W. Zillig.** 1991. Complete nucleotide sequence of the virus SSV1 of the archaebacterium *Sulfolobus shibatae.* Virology **185:**242–250.
23. **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, W. Brucker, V. Kumar, J. Kandasamy, L. Keenan, S. Bardarov, J. Kriakov, J. G. Lawrence, W. R. Jacobs, Jr., R. W. Hendrix, and G. F. Hatfull.** 2003. Origins of highly mosaic mycobacteriophage genomes. Cell **113:**171–182.
24. **Peng, X., I. Holz, W. Zillig, R. A. Garrett, and Q. She.** 2000. Evolution of the family of pRN plasmids and their integrase-mediated insertion into the chromosome of the crenarchaeon *Sulfolobus solfataricus.* J. Mol. Biol. **303:** 449–454.
25. **Reiter, W.-D., and P. Palm.** 1990. Identification and characterization of a defective SSV1 genome integrated into a tRNA gene in the archaebacterium *Sulfolobus* sp. B12. Mol. Gen. Genet. **221:**65–71.
26. **Reiter, W.-D., P. Palm, A. Henschen, F. Lottspeich, W. Zillig, and B. Grampp.** 1987. Identification and characterization of the genes encoding three structural proteins of the *Sulfolobus* virus-like particle SSV1. Mol. Gen. Genet. **206:**144–153.
27. **Reiter, W.-D., P. Palm, and S. Yeats.** 1989. Transfer RNA genes frequently serve as integration sites for prokaryotic genetic elements. Nucleic Acids Res. **17:**1907–1914.
28. **Reiter, W.-D., P. Palm, S. Yeats, and W. Zillig.** 1987. Gene expression in archaebacteria: physical mapping of constitutive and UV-inducible transcripts from the *Sulfolobus* virus-like particle SSV1. Mol. Gen. Genet. **209:** 270–275.
29. **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.
30. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
31. **Schleper, C., K. Kubo, and W. Zillig.** 1992. The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of transfection with viral DNA. Proc. Natl. Acad. Sci. USA **89:**7645–7649.
32. **Serre, M. C., C. Letzelter, J. R. Garel, and M. Duguet.** 2002. Cleavage properties of an archaeal site-specific recombinase, the SSV1 integrase. J. Biol. Chem. **277:**16758–16767.
33. **She, Q., R. K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M. J. Awayez, C. C.-Y. Chan-Weiher, I. G. Clausen, B. A. Curtis, A. D. Moors, G. Erauso, C. Fletcher, P. M. K. Gordon, I. H.-D. Jong, A. C. Jeffries, C. J. Kozera, N. Medina, X. Peng, H. P. Thi-Ngoc, P. Redder, M. E. Schenk, C. Theriault, N. Tolstrup, R. L. Charlebois, W. F. Doolittle, M. Duguet, T. Gaasterland, R. A. Garrett, M. A. Ragan, C. W. Sensen, and J. V. D. Oost.** 2001. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc. Natl. Acad. Sci. USA **98**:7835–7840.
34. **Stedman, K. M., C. Schleper, E. Rumpf, and W. Zillig.** 1999. Genetic requirements for the function of the archaeal virus SSV1 in *Sulfolobus solfataricus*: construction and testing of viral shuttle vectors. Genetics **152:**1397–1405.
35. **Stedman, K. M., Q. She, H. Phan, H. P. Arnold, I. Holz, R. A. Garrett, and W. Zillig.** 2003. Relationships between fuselloviruses infecting the extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2. Res. Microbiol. **154:** 295–302.
36. **Suzek, B. E., M. D. Ermolaeva, M. Schreiber, and S. L. Salzberg.** 2001. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics **17:**1123–1130.
37. **Swofford, D.** 2002. PAUP*4.0 phylogenetic analysis using parsimony (*and other methods), version 4 beta, 10th ed. Sinauer Associates, Sunderlund, Mass.
38. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:**4673–4680.
39. **Whitaker, R. J., G. W. Grogan, and J. W. Taylor.** 24 July 2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science 10.1126/science.1086909.
40. **Wood, A. G., W. B. Whitman, and J. Konisky.** 1989. Isolation and characterization of an archaebacterial viruslike particle from *Methanococcus voltae* A3. J. Bacteriol. **171:**93–98.
41. **Yeats, S., P. McWilliam, and W. Zillig.** 1982. A plasmid in the archaebacterium *Sulfolobus acidocaldarius.* EMBO J. **1:**1035–1038.
42. **Zillig, W., H. P. Arnold, I. Holz, D. Prangishvili, A. Schweier, K. Stedman, Q. She, H. Phan, R. Garrett, and J. K. Kristjansson.** 1998. Genetic elements in the extremely thermophilic archaeon *Sulfolobus.* Extremophiles **2:**131–140.
43. **Zillig, W., A. Kletzin, C. Schleper, I. Holz, D. Janekovic, J. Hain, M. Lanzendorfer, and J. K. Kristjansson.** 1994. Screening for *Sulfolobales*, their plasmids, and their viruses in Icelandic solfataras. Syst. Appl. Microbiol. **16:**609–628.
44. **Zillig, W., K. O. Stetter, S. Wunderl, W. Schulz, H. Preiss, and H. Scholz.** 1980. The Sulfolobus-"Caldariella" group: taxonomy on the basis of the structure of DNA-dependent RNA polymerases. Arch. Microbiol. **125**:259–269.