Supporting Information

Mochizuki et al. 10.1073/pnas.1203668109

SI Text

The functional complexity of the *Aeropyrum* coil-shaped virus (ACV) genome is far greater than that observed in other single-stranded (ss) DNA viruses. For example, ssDNA viruses have not been previously reported to encode genes for carbohydrate metabolism enzymes. ACV encodes two glycosyltransferases (ORF19 and -38), both of which are predicted to be specific for mannose (Table S1). Interestingly, adjacent to ORF38 (Fig. 5) is a gene for GDP-mannose 4,6-dehydratase (GMD; ORF39). Homologs of the two ACV glycosyltransferases are encoded by crenarchaeal filamentous viruses of the newly proposed order Ligamenvirales (1), which are known to glycosylate the major structural proteins of their virions (2). By contrast, the closest viral homologs of ORF39 are encoded by algae-infecting phycodnaviruses (3), in which GMD has indeed been shown to be active and capable of converting GDP-D-mannose into GDP-D-rhamnose (single handedly) and GDP-L-fucose (together with GDP-4-keto-6-deoxy-D-mannose epimerase/reductase). Notably, both fucose and rhamnose are components of the glycans attached to the major capsid protein of the phycodnavirus *Paramecium bursaria Chlorella* virus type 1 (4). It is therefore likely that the three carbohydrate metabolism enzymes encoded by ACV are similarly involved in the glycosylation of viral or/and cellular proteins. Another protein not typical for ssDNA viruses is a serine protease encoded by ORF25. Although the target of this protease cannot be inferred in silico, proteolytic cleavage is known to play critical roles at various stages in the reproduction of diverse RNA and DNA viruses, including during the coordinated assembly and maturation of virions (5–7).

Also unexpected for ssDNA viruses are the two thioredoxin-like genes (ORFs 12 and 13) of ACV (Fig. 5). Thioredoxins are small ubiquitous disulphide-containing redox proteins present in organisms from all three cellular domains (8). Certain complex double-stranded (ds) DNA viruses also encode thioredoxin-like proteins. In the context of the viral life cycle, these proteins were shown to perform a wide variety of functions, including direct oxidation of the thiols of vaccinia virus structural proteins (9) and scavenging of host nucleotides by reducing the myovirus T4-encoded ribonucleotide reductase (10). Alternatively, some viruses rely on the host-encoded thioredoxin. For example, cellular thioredoxin acts as a processivity factor on the T7 DNA polymerase (11). Importantly, host thioredoxin has been shown to be indispensable in the assembly of filamentous ssDNA inoviruses—specifically during the coating of the ssDNA with the major capsid protein and the concomitant removal of the ssDNA-binding protein (12). The products of ORFs 12 and 13 might also play an important role during virion assembly or some other step of the ACV infection cycle.

Viruses infecting Crenarchaeota have been previously found to encode a large number of DNA-binding proteins with ribbon-helix-helix (RHH) and winged helix-turn-helix (wHTH) motifs (13, 14), which are likely to be involved in transcription regulation (15). ACV encodes six such proteins (Fig. 5), suggesting that transcription of viral genes might be tightly regulated. Interestingly, ORF16 encodes a zinc ribbon protein that is similar in size and sequence to the RpoP subunit of the archaeal RNA polymerase (Fig. S7). Archaeal RpoP promotes DNA melting and stabilizes the open complex. It is therefore crucial during transcription initiation (16). It is tempting to speculate that the product of ORF16 is responsible for recruiting the cellular RNA polymerase to the viral template, similar to the way in which certain bacterial viruses encode their own σ-factors (17).

All known ssDNA viruses replicate—or are believed to replicate—their genomes via a rolling-circle (RCR; for circular genomes) or rolling-hairpin (for linear genomes) mechanism (18). This type of genome replication is initiated by a virus-encoded RCR Rep protein that nicks the dsDNA at a specific sequence and becomes covalently attached to the 5′ phosphate through a tyrosine residue present in its active site; the released free 3′ end then acts as a primer for DNA synthesis. The same Rep protein also catalyzes religation of the newly synthesized genomes once replication is complete. Without exception, all described ssDNA viruses encode related Rep proteins that share a set of conserved motifs (19). Surprisingly, we were not able to identify a potential ACV Rep candidate sharing significant sequence homology or conserved motifs with known RCR Rep proteins. Notably, however, the reaction catalyzed by tyrosine recombinases in many ways resembles that performed by RCR Reps. Enzymes from both families possess an active-site tyrosine that nicks one of the strands in the dsDNA, forming a covalent adduct (5′- and 3′-phosphotyrosyl bonds for RCR Reps and Tyr recombinases, respectively). In a subsequent step, the terminal hydroxyl group attacks the phosphotyrosine linkage to release the enzyme and regenerate duplex DNA. We hypothesize that the ACV ORF33 product, which is a divergent member of the tyrosine recombinase family, might be involved in ACV genome replication. If so, ORF33 should have evolved toward performing a reaction that deviates from that typical of other members in this family of recombinases. Future studies will reveal whether this is indeed the case.

SI Materials and Methods

Discovery and Isolation of ACV. In the course of a study of viruses infecting *Aeropyrum pernix*, the most extreme aerobic hyperthermophile, which grows optimally at 90–95 °C (20), we established and analyzed enrichment cultures from samples collected at the coastal Yamagawa Hot Spring in Kagoshima, Japan (21). Three types of particles were shown to represent virions: the *A. pernix* bacilliform virus 1 (21), *A. pernix* spindle-shaped virus 1, and *A. pernix* ovoid virus 1 (22). In addition to these viruses, we observed short, rod-shaped particles of uniform appearance and size in some enrichment cultures of *A. pernix*. The particles represented virions of a new virus, termed *Aeropyrum* coil-shaped virus or ACV. Occasionally, ACV virions made up ~95% of the particles present. They were collected from a cell-free growth culture by PEG precipitation or ultracentrifugation using a Beckman SW60 rotor at 48,000 × *g* for 12 h. Shorter times of centrifugation in the SW60 rotor at 48,000 × *g*, e.g., 3 h, were also sufficient to pellet the virions. The isopycnic gradient centrifugation of the pelleted material in caesium chloride yielded a sharp white opalescent band with a buoyant density of about 1.28 g·cm^{-3}.

For isolation of the viral hosts, 20 single strains were colony-purified from the enrichment culture. Colonies were formed on 0.8% Gelrite (CPKelco) plates containing 80% medium 3ST (21) supplemented with 0.12% tryptone and 0.08% yeast extract at 90 °C. All 20 isolates were strains of *A. pernix*, as indicated by their 16S rRNA gene sequences. In addition, *Aeropyrum camini* (23) (DSM 16960) and 40 single isolates of *A. pernix* were selected from the laboratory collection. All 61 strains of *Aeropyrum* were analyzed for their susceptibility to the infection with the ACV preparation. The virus was added to individually growing cultures of the 61 strains, and viral replication was verified by transmission electron microscopy observations. ACV was unable to replicate in any of the tested strains. Thus, for further ACV analyses, the original enrichment culture was maintained as a source of the virus.

Sequence Analysis. The ACV genome analysis was performed using the CLC Genomics Workbench software package (CLC Bio, Inc.). Translated ORFs were used as queries to search for sequence homologs in the nonredundant protein database at the National Center for Biotechnology Information using BLASTP (24) with an upper threshold *E*-value of 1*e*-5. Conserved protein domains were identified using CD-search (25). Searches for distant homologs were performed using HHpred (26) and FFAS03 (27). Transmembrane domains were predicted using TMHMM (http://www.cbs.dtu.dk/services/TMHMM/). Coiled-coil regions were identified using COILS (http://www.ch.embnet.org/software/COILS_form.html). The multiple sequence alignments were constructed using MUSCLE (28) or PROMALS3D (29).

1. Prangishvili D, Krupovic M (2012) A new proposed taxon for double-stranded DNA viruses, the order "Ligamenvirales." *Arch Virol* 157:791–795.
2. Vestergaard G, et al. (2005) A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus Acidianus. *Virology* 336:83–92.
3. Van Etten JL, Gurnon JR, Yanai-Balser GM, Dunigan DD, Graves MV (2010) Chlorella viruses encode most, if not all, of the machinery to glycosylate their glycoproteins independent of the endoplasmic reticulum and Golgi. *Biochim Biophys Acta* 1800: 152–159.
4. Tonetti M, et al. (2003) Paramecium bursaria Chlorella virus 1 encodes two enzymes involved in the biosynthesis of GDP-L-fucose and GDP-D-rhamnose. *J Biol Chem* 278: 21559–21565.
5. Hellen CU, Wimmer E (1992) The role of proteolytic processing in the morphogenesis of virus particles. *Experientia* 48:201–215.
6. Ganser-Pornillos BK, Yeager M, Pornillos O (2012) Assembly and architecture of HIV. *Adv Exp Med Biol* 726:441–465.
7. Steven AC, Heymann JB, Cheng N, Trus BL, Conway JF (2005) Virus maturation: Dynamics and mechanism of a stabilizing structural transition that leads to infectivity. *Curr Opin Struct Biol* 15:227–236.
8. Holmgren A (1985) Thioredoxin. *Annu Rev Biochem* 54:237–271.
9. Senkevich TG, White CL, Koonin EV, Moss B (2002) Complete pathway for protein disulfide bond formation encoded by poxviruses. *Proc Natl Acad Sci USA* 99:6667–6672.
10. Söderberg BO, Sjöberg BM, Sonnerstam U, Brändén CI (1978) Three-dimensional structure of thioredoxin induced by bacteriophage T4. *Proc Natl Acad Sci USA* 75: 5827–5830.
11. Huber HE, Russel M, Model P, Richardson CC (1986) Interaction of mutant thioredoxins of Escherichia coli with the gene 5 protein of phage T7. The redox capacity of thioredoxin is not required for stimulation of DNA polymerase activity. *J Biol Chem* 261: 15006–15012.
12. Russel M (1995) Moving through the membrane with filamentous phages. *Trends Microbiol* 3:223–228.
13. Prangishvili D, Garrett RA, Koonin EV (2006) Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life. *Virus Res* 117:52–67.
14. Krupovic M, White MF, Forterre P, Prangishvili D (2012) Postcards from the edge: Structural genomics of archaeal viruses. *Adv Virus Res* 82:33–62.
15. Guillière F, et al. (2009) Structure, function, and targets of the transcriptional regulator SvtR from the hyperthermophilic archaeal virus SIRV1. *J Biol Chem* 284: 22222–22237.
16. Reich C, et al. (2009) The archaeal RNA polymerase subunit P and the eukaryotic polymerase subunit Rpb12 are interchangeable in vivo and in vitro. *Mol Microbiol* 71: 989–1002.
17. Nechaev S, Severinov K (2003) Bacteriophage-induced modifications of host RNA polymerase. *Annu Rev Microbiol* 57:301–322.
18. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (2011) *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses* (Elsevier Academic Press, San Diego).
19. Ilyina TV, Koonin EV (1992) Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res* 20:3279–3285.
20. Sako Y, et al. (1996) Aeropyrum pernix gen. nov., sp. nov., a novel aerobic hyperthermophilic archaeon growing at temperatures up to 100 degrees C. *Int J Syst Bacteriol* 46:1070–1077.
21. Mochizuki T, et al. (2010) Diversity of viruses of the hyperthermophilic archaeal genus Aeropyrum, and isolation of the Aeropyrum pernix bacilliform virus 1, APBV1, the first representative of the family Clavaviridae. *Virology* 402: 347–354.
22. Mochizuki T, Sako Y, Prangishvili D (2011) Provirus induction in hyperthermophilic archaea: Characterization of Aeropyrum pernix spindle-shaped virus 1 and Aeropyrum pernix ovoid virus 1. *J Bacteriol* 193:5412–5419.
23. Nakagawa S, Takai K, Horikoshi K, Sako Y (2004) Aeropyrum camini sp. nov., a strictly aerobic, hyperthermophilic archaeon from a deep-sea hydrothermal vent chimney. *Int J Syst Evol Microbiol* 54:329–335.
24. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
25. Marchler-Bauer A, Bryant SH (2004) CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res* 32(Web Server issue):W327–331.
26. Söding J (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21:951–960.
27. Friedberg I, Jambon M, Godzik A (2006) New avenues in protein function prediction. *Protein Sci* 15:1527–1529.
28. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
29. Pei J, Tang M, Grishin NV (2008) PROMALS3D web server for accurate multiple protein sequence and structure alignments. *Nucleic Acids Res* 36(Web Server issue): W30–W34.

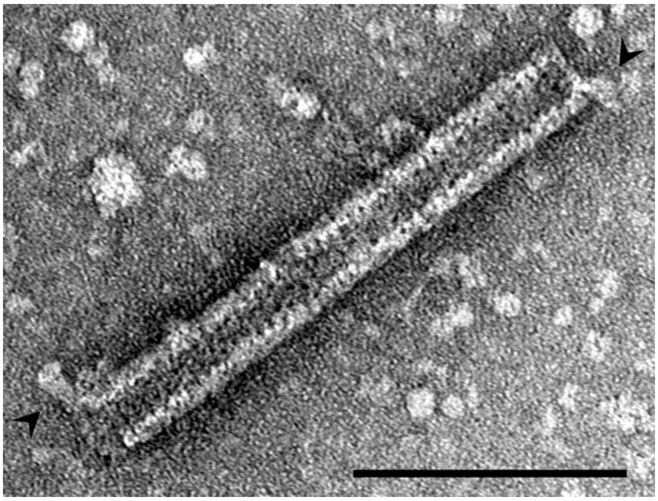


Fig. S1. Transmission electron micrograph of ACV virion with terminal appendages at opposite sides of the virion faces. Appendages are indicated by arrows. Sample was negatively stained with 2% (wt/vol) uranyl acetate. (Scale bar, 100 nm.)

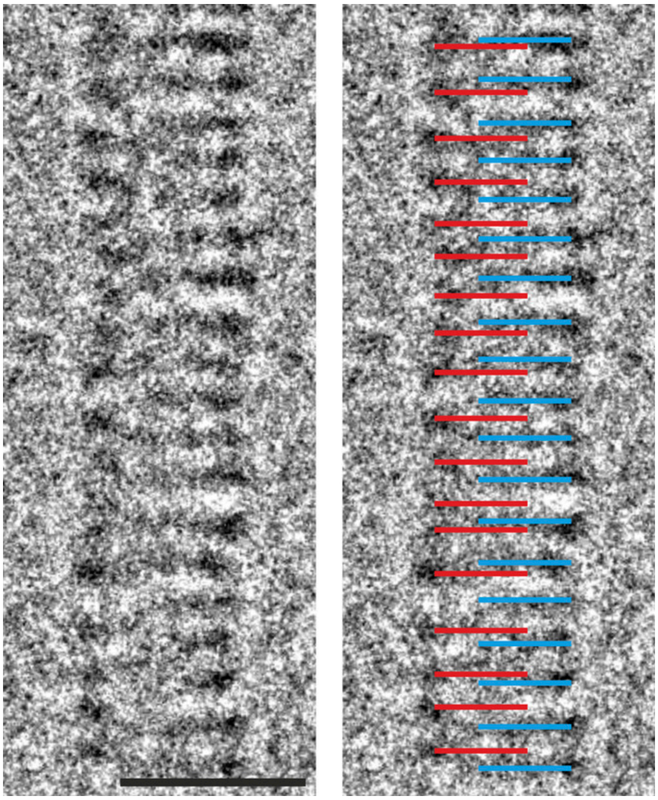


Fig. S2. Helical arrangement of the ACV virion. A fragment of a cryo-EM image of ACV virion. The staggered positions of higher density regions on the two sides of the fragment, highlighted with alternating red and blue lines, support the helifcal arrangement of the virion (see the main text for details). (Scale bar, 25 nm.)

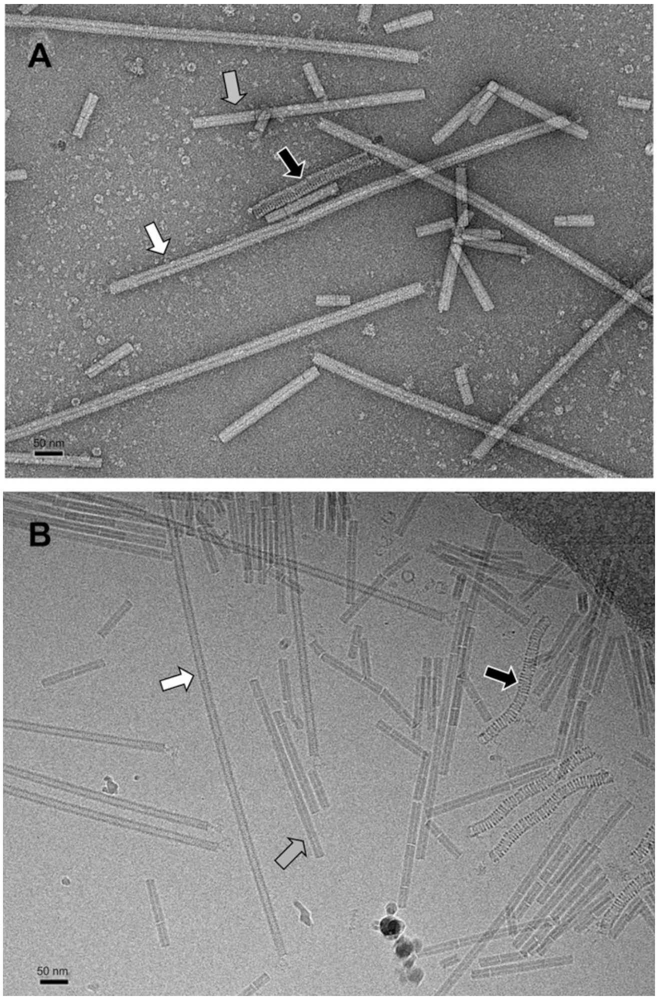


Fig. S3. Comparison of ACV to other helical viruses. (*A*) Virions of ACV (black arrows), tobacco mosaic virus (gray arrows), and *Sulfolobus islandicus* rod-shaped virus 2 (white arrows), negatively stained with 2% (wt/vol) uranyl acetate. (*B*) The same three types of virions as in *A*, embedded in vitreous ice. A certain proportion of tobacco mosaic virus particles appears broken in both *A* and *B*.

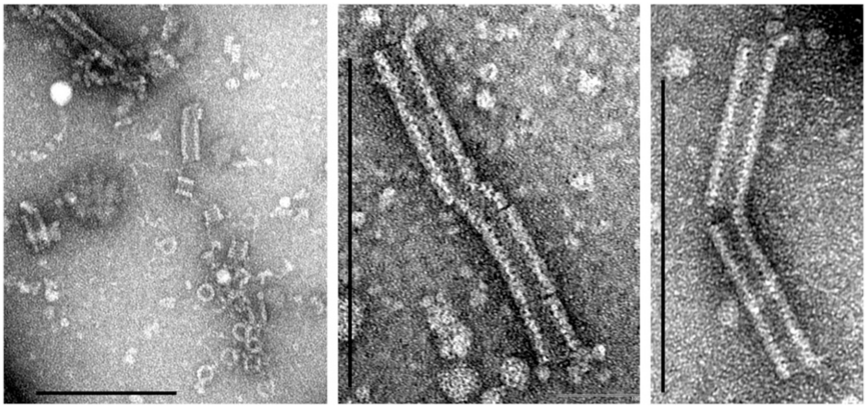


Fig. S4. Transmission electron micrograph of disrupted virions of ACV. Samples were negatively stained with 2% (wt/vol) uranyl acetate. (Scale bars, 200 nm.)

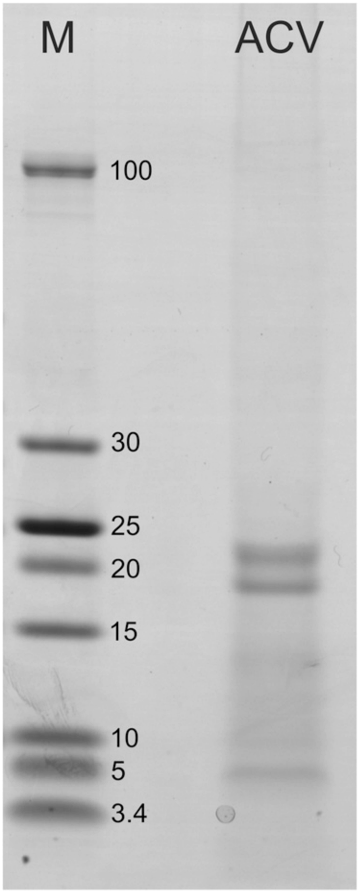


Fig. S5. Electrophoregram of SDS/PAGE of purified ACV virions. Molecular masses of standard proteins are indicated (M).

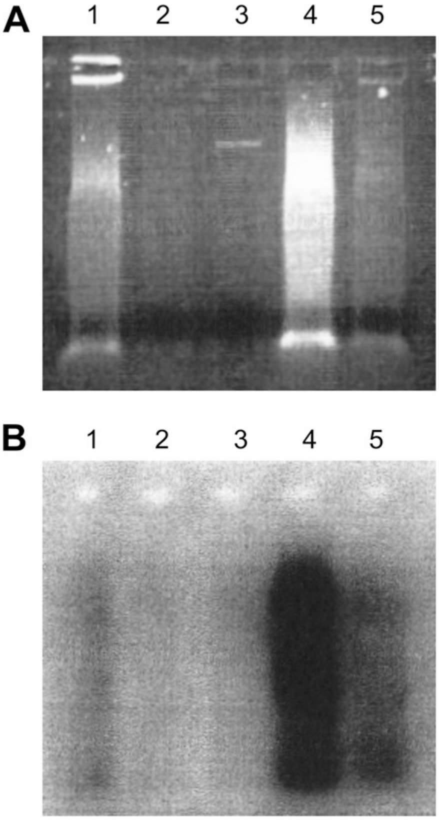
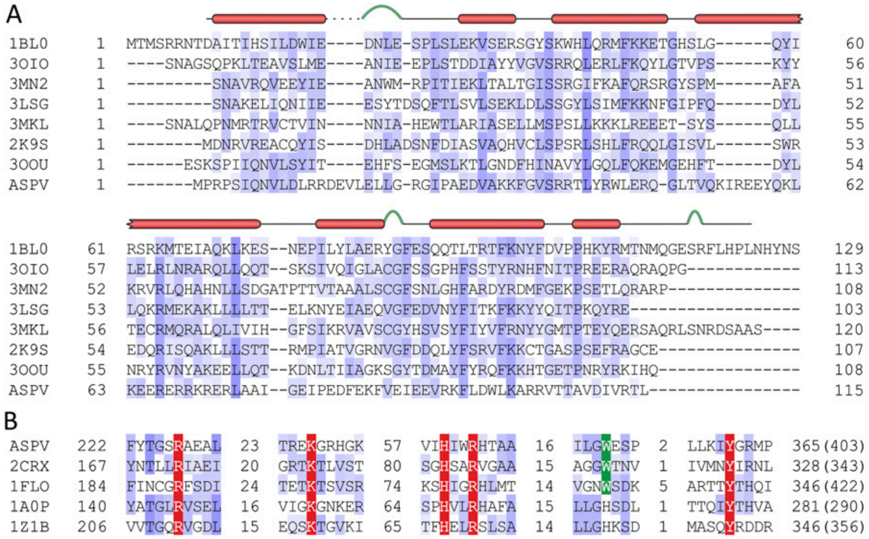


Fig. S6. Replication of ACV DNA and control DNAs by Klenow enzyme. Random hexamers (N6) were used as primers for replication. Primer elongation was verified by electrophoresis of products on 0.9% agarose gel. (*A*) Staining with ethidium bromide. (*B*) Autoradiogramm of DIG-labeled DNA. The following DNAs were used as templates: lane 1, 14 ng of M13 ssDNA; lane 2, 14 ng of M13 dsDNA; lane 3, 25 ng of linearized plasmid pBR328 dsDNA (provided with the kit); lane 4, 25 ng of linearized plasmid pBR328 dsDNA, heat-denatured before elongation; and lane 5, 14 ng of APSV viral DNA.



A

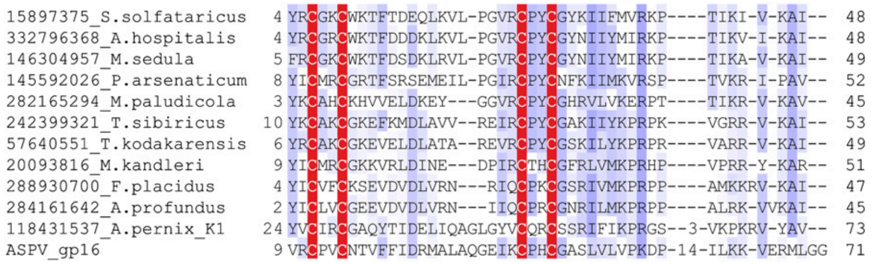
```
1BL0    1  MTMSRRNTDAITIHSILDWIE----DNLE-SPLSLEKVSERSGYSKWHLQRMFKKETGHSLG------QYI   60
3OIO    1  ----SNAGSQPKLTEAVSLME----ANIE-EPLSTDDIAYYVGVSRRQLERLFKQYLGTVPS------KYY   56
3MN2    1  ---------SNAVRQVEEYIE----ANWM-RPITIEKLTALTGISSRGIFKAFQRSRGYSPM------AFA   51
3LSG    1  ---------SNAKELIQNIIE----ESYTDSQFTLSVLSEKLDLSSGYLSIMFKKNFGIPFQ------DYL   52
3MKL    1  ----SNALQPNMRTRVCTVIN----NNIA-HEWTLARIASELLMSPSLLKKKLREEET-SYS------QLL   55
2K9S    1  --------MDNRVREACQYIS----DHLADSNFDIASVAQHVCLSPSRLSHLFRQQLGISVL------SWR   53
3OOU    1  ------ESKSPIIQNVLSYIT----EHFS-EGMSLKTLGNDFHINAVYLGQLFQKEMGEHFT------DYL   54
ASPV    1  -------MPRPSIQNVLDLRRDEVLELLG-RGIPAEDVAKKFGVSRRTLYRWLERQ-GLTVQKIREEYQKL   62

1BL0   61  RSRKMTEIAQKLKES--NEPILYLAERYGFESQQTLTRTFKNYFDVPPHKYRMTNMQGESRFLHPLNHYNS  129
3OIO   57  LELRLNRARQLLQQT--SKSIVQIGLACGFSSGPHFSSTYRNHFNITPREERAQRAQPG------------  113
3MN2   52  KRVRLQHAHNLLSDGATPTTVTAAALSCGFSNLGHFARDYRDMFGEKPSETLQRARP--------------  108
3LSG   53  LQKRMEKAKLLLLTT--ELKNYEIAEQVGFEDVNYFITKFKKYYQITPKQYRE------------------  103
3MKL   56  TECRMQRALQLIVIH--GFSIKRVAVSCGYHSVSYFIYVFRNYYGMTPTEYQERSAQRLSNRDSAAS----  120
2K9S   54  EDQRISQAKLLLSTT--RMPIATVGRNVGFDDQLYFSRVFKKCTGASPSEFRAGCE---------------  107
3OOU   55  NRYRVNYAKEELLQT--KDNLTIIAGKSGYTDMAYFYRQFKKHTGETPNRYRKIHQ---------------  108
ASPV   63  KEERERRKRERLAAI--GEIPEDFEKFVEIEEVRKFLDWLKARRVTTAVDIVRTL----------------  115
```

B

```
ASPV   222  FYTGSRAEAL  23  TREKGRHGK  57  VIHIWRHTAA  16  ILGWESP  2  LLKIYGRMP  365(403)
2CRX   167  YNTLLRIAEI  20  GRTKTLVST  80  SGHSARVGAA  15  AGGWTNV  1  IVMNYIRNL  328(343)
1FLO   184  FINCGRFSDI  24  TETKTSVSR  74  KSHIGRHLMT  14  VGNWSDK  5  ARTTYTHQI  346(422)
1A0P   140  YATGLRVSEL  16  VIGKGNKER  64  SPHVLRHAFA  15  LLGHSDL  1  TTQIYTHVA  281(290)
1Z1B   206  VVTGQRVGDL  15  EQSKTGVKI  65  TFHELRSLSA  14  LLGHKSD  1  MASQYRDDR  346(356)
```

Fig. S7. Multiple alignment of the ACV ORF33 gene product with related sequences. The alignment is colored according to the sequence conservation (BLOSUM62 matrix). (*A*) Alignment of the N-terminal domain of ORF33 with the selected helix-turn-helix proteins for which X-ray structures are available. The secondary structural elements determined from the X-ray structure of MarA from *Escherichia coli* are indicated above the alignment: α-helixes, red ellipsoids; turns, green bulges. The proteins are indicated with their Protein Data Bank (PDB) identification numbers: 1BL0, MarA from *E. coli*; 3OIO, AraC-like transcriptional regulator from *Chromobacterium violaceum*; 3MN2, AraC-like transcriptional regulator from *Rhodopseudomonas palustris* CGA009; 3LSG, yesN from *Fusobacterium nucleatum*; 3MKL, transcriptional regulator from *E. coli* K-12; 2K9S, AraC from *E. coli*; 3OOU, lin2118 from *Listeria innocua*. (*B*) Alignment of the C-terminal domain of ORF33 with the selected tyrosine recombinases. Only the regions encompassing the active-site residues (highlighted in red) are shown. The active-site tryptophan found in Cre and Flp recombinases, but not in the other proteins, are highlighted in green. Numbers between the motifs indicate the spacing between the corresponding motifs. The proteins are indicated with their PDB identification numbers: 2CRX, Cre recombinase from bacterial myovirus P1; 1FLO, Flp recombinase from *Saccharomyces cerevisiae*; 1A0P, XerD recombinase from *E. coli*; 1Z1B, integrase from siphovirus λ.



```
15897375_S.solfataricus    4 YRCGKCWKTFTDEQLKVL-PGVRCPYCGYKIIFMVRKP----TIKI-V-KAI--  48
332796368_A.hospitalis     4 YRCGRCWKTFDDDKLKVL-PGVRCPYCGYNIIYMIRKP----TIKV-I-KAI--  48
146304957_M.sedula         5 FRCGKCWKTFDSDKLRVL-PGVRCPYCGYNIIYMIRKP----TIKA-V-KAI--  49
145592026_P.arsenaticum    8 YICMRCGRTFSRSEMEIL-PGIRCPYCNFKIIMKVRSP----TVKR-I-PAV--  52
282165294_M.paludicola     3 YKCAHCKHVVELDKEY---GGVRCPYCGHRVLVKERPT----TIKR-V-KAV--  45
242399321_T.sibiricus     10 YKCAKCGKEFKMDLAVV--REIRCPYCGAKIIYKPRPK----VGRR-V-KAI--  53
57640551_T.kodakarensis    6 YRCAKCGKEVELDLATA--REVRCPYCGSKILYKPRPR----VARR-V-KAI--  49
20093816_M.kandleri        9 YICMRCGKKVRLDINE---DPIRCTHCGFRLVMKPRHP----VPRR-Y-KAR--  51
288930700_F.placidus       4 YICVFCKSEVDVDLVRN---RIQCPKCGSRIVMKPRPP----AMKKRV-KAI--  47
284161642_A.profundus      2 YICLVCGEEVDVDLVRN---IIQCPRCGNRILMKPRPP----ALRK-VVKAI--  45
118431537_A.pernix_K1     24 YVCIRCGAQYTIDELIQAGLGYVCQRCSSRIFIKPRGS--3-VKPKRV-YAV--  73
ASPV_gp16                  9 VRCPVCNTVFFIDRMALAQGEIKCPHCGASLVLVPKDP-14-ILKK-VERMLGG  71
```

Fig. S8. Multiple sequence alignment of the ACV ORF16 gene product with the RpoP subunits of archaeal DNA-dependent RNA polymerases. The four conserved cysteine residues are highlighted in red. The sequences are indicated with their GenBank accession numbers followed by the corresponding organism names. The alignment is colored according to the sequence conservation (BLOSUM62 matrix).

Table S1. Summary of predicted ORFs in the ACV genome

ORF	Length (aa)	TMDs	Function	HMM predictions	BLAST hit	Identity (%)	*E*-value
1	57						
2	78						
3	208						
4	78						
5	178	1 (15–37)			ACV ORF6	34/133 (26%)	2e-05
6	144				ACV ORF5	38/132 (29%)	2e-07
7	100						
8	43						
9	303	1 (239–261)					
10	261						
11	407						
12	154	1 (11–33)	Thioredoxin	Thioredoxin superfamily; thioredoxin (aa 38–145; hit to 2j23; HHpred Probab = 97.66).	*Singulisphaera acidiphila* DSM 18658 (ZP_09569153); *Listeria* phage A511 (YP_001468508)	30/93 (32%) 20/70 (29%)	7e-02 4e-05
13	124		Thioredoxin	Thioredoxin superfamily; theoredoxin (aa 9–124; hit to 3hz4; HHpred Probab = 99.94)	*Bacillus subtilis* (ZP_03598266); Mycobacterium phage Kugel (AER49994)	28/77 (36%) 15/59 (25%)	0.11 2e-04
14	47	1 (13–35)			ACV ORF15; ACV ORF44	14/34 (41%) 20/42 (48%)	7e-04 9e-08
15	40	1 (10–32)			ACV ORF44; ACV ORF14	19/41 (46%) 14/34 (41%)	9e-06 7e-04
16	72		Zinc ribbon protein	Zinc ribbon; archaeal rpoP-like protein (aa 11–39) HHpred: hit to 2waq_P; Probab = 97.5; FFAS03: hit to 2pmz_P, score −9.51	*Desulfovibrio alaskensis* G20 (YP_004849538) *Metallosphaera sedula* DSM 5348 (YP_001192273)	18/45 (40%) 14/45 (31%)	2e-03 3e-05
17	89	1 (62–84)		Coiled-coil domain protein; predicted SlyX homolog (aa 1–61; hit to 3efg; HHpred Probab = 96.75; FFAS03 score −11.5); CDD prediction: chromosome segregation protein SMC, primarily archaeal type (TIGR02169; 6.5e-03)	*Synechococcus* phage P60 (NP_570325)	24/53 (45%)	3e-04
18	67	2 (5–22, 37–59)					
19	**334**		**Glycosyltransferase**	**Glycosyltransferase (GT1 family; CDD: cd03801, PF00534; *E* = 1.6e-19); HHpred: hit to mannosyltransferase, 3okp; Probab = 100**	***Sulfolobus islandicus* filamentous virus (NP_445705);** ***Nakamurella multipartita* DSM 44233 (YP_003203461)**	**53/173 (31%)** **56/168 (33%)**	**3e-16** **4e-11**
20	52	1 (7–29)					
21	227	1 (194–216)					
22	102	2 (34–56, 71–93)					
23	85	2 (12–34, 49–71)					
24	54	1 (15–37)					
25	**326**		**Serine protease**	**Trypsin-like serine protease (aa 105–322; hit to 2sga; HHpred Probab = 99.43; FFAS03 score −36.8, hit to 2sfa)**	***Archaeoglobus veneficus* SNP6 (YP_004342149)**	**111/317 (35%)**	**3e-39**
26	143	1 (109–131)					
27	58						
28	37						
29	347		DNA binding (RHH)	DNA-binding protein; N-terminal RHH domain, central and C-terminal coiled-coil domains; aa 4–41: hit to RHH (2ay0), HHpred Probab = 96.95	*N*-terminal: *Sulfolobus tokodaii* strain 7 (NP_377238)	14/36 (39%)	3e-04
30	86						
31	69						
32	62	2 (7–26, 36–58)					

Table S1. Cont.

ORF	Length (aa)	TMDs	Function	HMM predictions	BLAST hit	Identity (%)	*E*-value
33	403		Recombinase	N-terminal (aa 1–115) AraC-like double HTH domain; hit to 3gbg, HHpred Probab = 99.1; aa 93–368, Cre recombinase/phage integrase, hit to 1xo0, HHpred Probab = 98.12; CDD: INT_REC_C (cd01182), *E* = 8.5e-05	*Bacillus* phage phBC6A51 (NP_852546; *Candidatus Nitrosoarchaeum koreensis* MY1 (ZP_08667533)	74/282 (26%) 95/407 (23%)	2e-06 5e-05
34	73			Coiled-coil domains			
35	134		DNA binding (RHH)	Coiled-coil domain; C-terminal (aa 74–126) RHH domain (PF03693); hit to ParD antitoxin protein (3kxe; HHpred Probab = 96.60, FFAS03 score −9.97)	Pyrobaculum sp. 1860 (AET34113)—similarity limited to RHH domain	18/34 (53%)	7e-06
36	43	1 (20–42)					
37	93						
38	**318**		**Glycosyltransferase**	**Glycosyltransferase (GT1 family; CDD: cd03801, PF00534; *E* = 1.2e-09); HHpred: hit to mannosyltransferase, 3okp; Probab = 99.98**	***Actinomyces urogenitalis* DSM 15434 (ZP_03928147); *Stygiolobus* rod-shaped virus (CAQ58459)**	**57/186 (31%)** **55/227 (24%)**	**4e-10** **9e-14**
39	**390**		**Carbohydrate modification (GMD)**	**Carbohydrate modification (GDP-mannose 4,6 dehydratase, extended (e) SDRs; CDD: cd05260, *E* = 7.1e-113); HHpred: hit to GDP-mannose 4,6-dehydratase, 2z1m, Probab = 100.00; FFAS03 score −113**	**Methanobacterium sp. AL-21 (YP_004290508); *Paramecium bursaria Chlorella* virus AR158 (YP_001498237)**	**163/397 (41%)** **134/382 (35%)**	**3e-87** **4e-58**
40	219				Highly divergent homolog of APHV ORF42 (48/209; 23%)		
41	199						
42	221						
43	80		DNA binding (wHTH)	wHTH domain (aa 9–79; hit to 1y0u, HHpred Probab = 94.78)			
44	65	1 (18–40)			ACV ORF14; ACV ORF15	20/42 (48%) 19/41 (46%)	9e-08 9e-06
45	90						
46	91						
47	86						
48	41						
49	111			Hit (aa 2–89) to hypothetical protein 2obb HHpred Probab = 93.77	*Halorhabdus utahensis* DSM 12940 (YP_003130625)	31/115 (27%)	3e-03
50	51						
51	45						
52	134						
53	94		DNA binding (RHH)	*N*-terminal RHH domain; hit (aa 1–42) to 2ay0, HHpred Probab = 99.58; C-terminal coiled-coil domain	Thermophilic uncultured bacterium (BAF45195); *Thermococcus barophilus* MP (YP_004072303)	28/85 (33%) 17/32 (53%)	1e-03 1e-04
54	119		DNA binding (wHTH)	wHTH	*Ferroglobus placidus* DSM 10642 (YP_003436490)	32/107 (30%)	5e-04
55	80		DNA binding (wHTH)	wHTH	*Aeropyrum pernix* K1 (NP_148656)	19/54 (35%)	2e-04
56	66						
57	73						

ACV gene products that share significant similarity (BLASTP, cutoff of *E* = 1e-05) with sequences in the nonredundant protein database are indicated by boldface type. TMD, Trans membrane domain; HMM, hidden Markov model; CDD, Conserved Domain Database.

Table S2. Families of viruses with ssDNA genomes

Family	Morphology	Virion size (nm)	Genome structure	Genome size	Strand	Host*
"Spiraviridae"[†]	Spring-shaped	230 × 25	Circular	24.9	+	A
"Pleolipoviridae"[†]	Pleiomorphic	44 × 55	Circular	7.0–8.0	+	A
Inoviridae (Inovirus)	Filamentous	700–3,500 × 7	Circular	5.8–12.4	+	B
Inoviridae (Plectrovirus)	Rod-shaped	200–400 × 15	Circular	4.5–8.2	+	B
Anelloviridae	Icosahedral	30	Circular	2–4	–	E
Circoviridae	Icosahedral	12–27	Circular	1.7–2.3	– or ±	E
Geminiviridae	Icosahedral	38 × 28	Segmented circular	2.5–3/segment	– or ±	E
Microviridae	Icosahedral	25–27	Circular	4.4–6.1	+	E
Nanoviridae	Icosahedral	18–20	Segmented circular	1/segment	+	E
Parvoviridae	Icosahedral	21–26	Linear	4–6.3	±	E

*A, Archaea; B, Bacteria; E, Eukarya.

[†] Proposed families.

Table S3. PCR primers used in DNA amplification experiments

Name	Sequence	Positions in viral genome
FR26RV-N	GCC GGA GCT CTG CAG ATA TCN NNN NN	—
FR20RV	GCC GGA GCT CTG CAG ATA TC	—
seq_1	GGC TTC TGC AAG GCT TAA TG	19,925–19,906
seq_2	CGC TTG AGA GTT GGC TTA GG	21,148–21,167
1F	CTT ACT ACT TCT TAA CCC AAA GGG AGT TAT	2,425–2,454
2F	GAT ACG ATA TTG TTA AGA GAC AGA TAT CCA	8,210–8,239
3F	CTA TAG AGG CTA CAC TAA TAG GGA CAC C	17,155–17,182
4R	TAG TTA ACA GAA TAT TGA GAA CCT CCA GTA	2,953–2,924
5R	ATC AGT CTC ACT ATT ACT TAC TAT GCC AAC	11,730–11,701
6R	GAT ATG AAC TGT TTA CCA TAC TCC TCA CT	17,853–17,825

—, absent in the ACV viral genome. The positions in the viral genome are determined in respect to the start codon of the ORF1.