Identification and characterization of a novel *Geobacillusthermoglucosidasius* bacteriophage, GVE3

Leonardo Joaquim van Zyl[1*],Falone Sunda[1],Mark Paul Taylor[2], Don Arthur Cowan[1,3], Marla Iris Trindade[1]

[1]Institute for Microbial Biotechnology and Metagenomics (IMBM), University of the Western Cape, Robert Sobukwe Road, Bellville, Cape Town, South Africa.

[2] TMO Renewables Limited, 40 Alan Turing Road, The Surrey Research Park, Guildford, Surrey, GU2 7YF, UK

[3]Centre for Microbial Ecology and Genomics, Department of Genetics, University of Pretoria, Pretoria 0002, South Africa

* Author to whom correspondence should be addressed; e-mail: vanzyllj@gmail.com; Tel.: +27 21 9592325

Email: Leonardo J van Zyl vanzyllj@gmail.com – FaloneSundafalone.Sunda@gmail.com – Mark P. Taylor marktaylorimbm@gmail.com - Marla I Tuffin prof.marlatt@gmail.com - Don A Cowan don.cowan@up.ac.za

Keywords: Bacteriophage, *Geobacillus*, *Siphoviridae*

Abstract

The study of extremophilicphages may reveal new phage families as well as different mechanisms of infection, propagation and lysis to those found in phages from temperate environments. We describe a novel siphovirus, GVE3, that infects the thermophile*Geobacillusthermoglucosidasius*. The genome size is 141298 bp(G+C 29.6%) making it the largest *Geobacillus*spp infecting phage known.GVE3 appears to be most closely related to the recently described *Bacillus anthracis* phage vB_BanS_Tsamsa, rather than*Geobacillus* infecting phages described thus far.Tetranucleotide usage deviation analysis supports this relationship, showing that the GVE3 genome sequence correlates best with *B. anthracis* and *Bacillus cereus* genome sequences, rather than *Geobacillus*spp genome sequences.

Introduction

The ubiquity of bacteriophages (phages) in nature and their impact on various trophic levels is widely appreciated [58; 76]. As phagesdirectlyaffect microbial communities that play a pivotal role in biogeochemical cycles, they in turnplay a role in alteringthose cycles[18; 33; 74]. Phages are also known to be prevalent in many extreme environments including soda lakes, terrestrial hot springs, deep sea hydrothermal vents, hot/cold deserts and hypersalinesystems, with some of the highest phage numbers being recorded in these habitats [40]. However, few studies have investigated the functional relationships between extremophiles and the phages that infect them, compared to the wealth of data that exist for phages and hosts in temperate environments.

Morphological and sequence based characterization of phages from many temperate environments has shown the predominance of tailed viruses (Caudovirales) with *Siphoviridae*, *Myoviridae* and *Podoviridae* most often recorded [3; 4; 68; 71].Morphological characterization of extremophilicphageshas led to the introduction of several new families including *Liptothrixviridae*, *Rudiviridae* and *Fuselloviridae*[6]. The study of extremophilicphages has also revealed new mechanisms for host lysis, as in the case of the deep-sea thermophilic bacteriophage GVE2 [17],and have demonstrated interactions between phage and host proteins which are unlike those normally observed for mesophilicphages[32].*Thermusthermophilus*phage φYS40 promotersare thought to beleaderless (i.e., contain no -10 or -35 elements), unlike those found in T4 and many other mesophilicphages which require phage- and host-encoded sigma factors for transcription [70].

It is therefore likely that the further study of phages infecting extremophiles will reveal new phage families, alternate strategies for infection or the "decision" between lysis and lysogeny,propagation,andwill shed further light on the behaviour and the role of host organisms in their natural environments[44, 57; 62]. Extremophilicphagesmay also provide a source of novel enzymes, adapted to extreme conditions, and serve as the basis for the development of genetic systems by providing strong regulatable promoters and as vehicles for the introduction of large DNA segments into bacterial hosts for which no genetic tools currently exist [53; 59].

Geobacillusthermoglucosidasius isa Gram positive thermophile which has been isolated from soil, oil fields, compost heaps, deep sea sediment and hot springs [54; 67].This promising "platform" organism is capable of producing a range of useful metabolites including ethanol, isobutanol and polylactic acid [19; 42; 79;http://tinyurl.com/po6a52q].Several *Geobacillus*speciesphages have been described (GVE1, GVE2, GBSV1, GBK2, DE6 and ϕOH2), sequenced and studied[20; 33; 43; 45; 72;73; 82; 83],although none infecting *G.*

*thermoglucosidasius*have been reported. Here we describe the first phage (GVE3) known to specifically infect *G. thermoglucosidasius*.

Materials and Methods

Media, bacterial strains and plasmids

G. thermoglucosidasius strains were cultured intryptone glycerol pyruvate(TGP) medium. One liter of TGP broth contains 17 g tryptone, 3 g Soy peptone, 2.5 g K_2HPO_4 and 5 g NaCl. The pH was adjusted to 7.3 before autoclaving, after which 4 g Na-pyruvate and 4 mL glycerol (filter sterilized) were added. For solid media, 15 g/L agar was added before autoclaving. TGP was used for general maintenance of cultures. Cultures were incubated 60°C with vigorous aeration.

DNA manipulations and sequencing

Plasmid preparations, restriction endonuclease digestions, gel electrophoresis and ligations were performed using standard methods or following the manufacturers' recommendations. Total DNA from all bacterial strains was prepared as described [34]. Phage DNA was prepared by first preparing a phage lysate from 1L of culture as described below. The phage was pelleted by centrifugation at 13000 x g for 30min after addition of PEG8000 (7.5ml of 20% PEG8000 per 30ml lysate) and incubation at 4°C overnight. The pellet was resuspended in 1ml SM buffer (5.8 g/L NaCl, 1.2 g/L MgSo4, 50 mL 1M Tris-HCl, pH 7.5, 0.1 g/L Gelatin). The suspension was treated withDNaseI and RNaseA (Fermentas; final concentration of 0.1µg/ml) at 37°C for 1 hour (DNaseI). The presence of contaminating bacterial DNA was checked by amplifying the 16SrRNAgene. The suspension was treated with ProteinaseK (Fermentas - final concentration 1ug/ml) at 55°C for 2 hours, before addition of 70µl 20% (wt/vol) SDS and incubation at 37°C for 1 hour. An equal volume of phenol:chloroform:isoamylalcohol (P:C:I; 25:24:1) was added, the sample centrifuged (15ml Sterillin tube, Eppendorf 5810R centrifuge, 5000 RPM for 10min) to separate the phases and the top aqueous phase removed to a fresh tube. A second P:C:I extraction was performed. An equal volume of C:I (24:1) was added to the supernatant and re-centrifuged. The top phase was removed to a fresh tube and a tenth volume of 3M sodium acetate (pH5.2) and two volumes of 100% ethanol were added. This was incubated at 4°C to precipitate overnight. The sample was centrifuged at 13000 RPM for ten minutes to pellet the DNA, and the pellet resuspended in 40ul of TE buffer. The phage DNA

was electrophoresed on a 1% low melting point agarose gel, excised and purified from the gel using standard agarase (Fermentas) treatment. The pellet was resuspended in 40µl TE buffer. The quality and integrity of the DNA was checked using a Bioanalyzer prior to library preparation. Sanger DNA sequencing was performed using an ABI Prism 377 automated DNA sequencer (University of Stellenbosch Central Analytical Facility) while Next Generation sequencing was performed using either a Roche GS Junior with a LibL library preparation kit, or an IlluminaMiSeqwith the Nextera XT 150bp library kit (Illumina). The raw reads were trimmed and de-multiplexed at the sequencing facility (the University of the Western CapeNext Generation Sequencing facility), resulting in two (2 x 150) paired fastq files. Sequences were analyzed with DNAMAN (version 4.1, LynnonBioSoft), Newbler (Roche) or CLC Genomics Workbench version 6.5 (CLC Bio). Open reading frames were predicted using the built in tools in the CLC Genomics workbench and confirmed by BLASTp against the NCBInr database. Smaller ORF's not identified by the software were assigned through manual translation of DNA sequences and BLASTp analysis of putative ORF's[5]. The complete genome sequence of *G. thermoglucosidasius* bacteriophage GVE3 is available on the GenBank database under accession no. KP144388. RAST [http://rast.nmpdr.org/; 7] and PHAST (http://phast.wishartlab.com/; 86) was used to identify closely related phages. RADAR was used to identify protein repeat regions (http://www.ebi.ac.uk/Tools/pfa/radar/). Direct repeats were identified using REPFIND [http://zlab.bu.edu/repfind/form.html;9] with a 15bp minimum repeat length. Inverted repeats were identified using UGENE (http://ugene.unipro.ru/) with a 20bp minimum and 80% similarity as search parameters. tRNA genes were predicted using the tRNAscan-SE program [http://lowelab.ucsc.edu/tRNAscan-SE/; 46] and ARAGORN [http://mbio-serv2.mbioekol.lu.se/ARAGORN/; 39]. Transmembrane regions were predicted using the TMHMM server v2.0 [http://www.cbs.dtu.dk/services/TMHMM/; 36]. Intron prediction was done using the RNAweasel server [http://megasun.bch.umontreal.ca/RNAweasel/; 38]

Polymerase chain reaction

Polymerase chain reaction (PCR) was performed using Phusion DNA polymerase (New England Biolabs™). Generally, 50 ngof DNA was used in a 50 µl reaction volume containing 2 mM $MgCl_2$, 0.125 µM of each primer, 0.2 mM of each deoxynucleoside triphosphate, and 1 U DNA polymerase.Reactions were carried out in a Biorad T-100 thermocycler, with an initial denaturation at 98 °C for 3 min, followed by 30 cycles of denaturation (30 s at 98 °C), annealing (30 s), and variable elongation timesat 72 °C as required.

Phage purification, maintenance and characterization

Phage lysates were prepared by culturing *G. thermoglucosidasius* to an $OD_{600nm}$of 0.4 and addition of phage particles at a multiplicity of infection (MOI) of 10. Infected cultures were incubated until complete culture lysis was observed. 1/10 volume of chloroform was added to lyse residual bacterial cells and release bacteriophage. Cell debris and chloroform were removed by centrifugation (5000RPM for 10min)and the supernatant was recovered as the phage stock.

The lysate was diluted in TGP broth and used in standard overlay plaque assays with sloppy agar (0.3% wt/vol agar). Single plaques from these assays were picked using a cut pipette tip to stab into the agar and lift the plaques from the plate. Plaques were crushed and suspended in 1ml TGP broth then used in subsequent rounds of plaque assays. Three rounds of plaque purification were performed and the purified phages used in all subsequent studies.

Mass spectrometry

Samples were precipitated using 5 volumes of ice cold acetone and incubated overnight at -20°C. Precipitates were pelleted by centrifugation at 12 000 X g for 10 min. Supernatants were carefully removed and pellets air-dried prior to dissolution in 100mM triethylammonium bicarbonate (TEAB) and determination of protein concentrations (A_{280nm}). 100µg aliquots of solubilized proteins were reduced with 5 mMtris-carboxyethyl phosphine (TCEP; Fluka) for 30 minutes at room temperature. Cysteine residues were methylated with 10 mM methane methylthiosulfonate (MMTS; Sigma) for 15 minutes at room temperature. After methylation, samples were diluted to 95 µL with 50mM TEAB before the addition of 5 µL trypsin (Promega) at 1mg/mL.Sample were incubated at 37°C overnight, dried and resuspended in 30 µL 2% acetonitrile:water/0.05% TFA.

Residual digest reagents were removed using an in-house manufactured C18 stage tip. The samples were loaded onto the stage tip after activating the C18 membrane with 30 µL methanol (Sigma) and equilibration with 30 µL 2% acetonitrile:water/ 0.05% TFA. The bound sample was washed with 30 µL 2% acetonitrile:water/ 0.05% TFA before elution with 30 µL 50% acetonitrile:water/ 0.05% TFA. The eluate was evaporated to dryness. The dried peptides were dissolved in 2% acetonitrile:water; 0.1% TFA for LC-MS analysis. Liquid chromatography was performed on a Thermo Scientific Ultimate 3000 RSLC equipped with a 2cmx100 µm C18 trap column and a 25cmx75 µm Pepmap C18 analytical column. The solvent system employed was loading: 2% acetonitrile:water/ 0.1 TFA; Solvent A: 2% acetonitrile:water/ 0.1% TFA and Solvent B: 80% acetonitrile:water.

The samples were loaded onto the trap column using loading solvent at a flow rate of 5 µL/min from a temperaturecontrolled autosampler set at 7°C. Loading was performed for 10 min before the sample was eluted onto the analytical column. The gradient was generated at 300nL/min as follows 0-4min in 2%A; 4-6min 6%A; 6-95min 6-35%A (Chromeleon non-linear gradient 6); 95-100min 35-50%A. Chromatography was performed at 50°C and the outflow delivered to the mass spectrometer through a stainless steel nano-bore emitter. Mass spectrometry was performed on a Thermo Scientific Fusion mass spectrometer. Data was acquired in positive mode using a Nanospray Flex (Thermo Scientific) nano-ESI source with spray voltage set to 1.7 kV and ion transfer tube temperature set to 300°C. MS1 scans were recorded in the Orbitrap mass analyser set to 12 000 resolution over the scan range m/z = 350-1650 with a fill time of 50 ms or until adaptive gain control (AGC) target of 4e5 were reached. Ion filter criteria were set to mono-isotopic precursors only with charge state 2-6 and dynamic exclusion of 1 over 40s with mass tolerance of 10 ppm. Precursor selection was performed in Top Speed data dependent mode with the most intense precursor selected first with a cut off intensity higher than 50 000.Precursor selection was performed using the quadrupole mass analyser with an isolation window of m/z =1.5 prior to HCD fragmentation. HCD collision energy was set to 35%. Detection was performed in the ion trap mass analyser with ion injection time of 40ms or until an AGC target of 1e4 was reached. The raw files generated by the mass spectrometer were imported into Proteome Discoverer v1.4 (Thermo Scientific) and processed using the Sequest HT. Database interrogation was performed against GVE3 predicted ORF sequences with trypsin cleavage allowing for 2 missed cleavages. Precursor mass tolerance was set to 10ppm and fragment mass tolerance set to 0.8 Da. Deamidation (NQ) and oxidation (M) was allowed as dynamic modifications and thiomethyl of C as static modification.

Electron microscopy

Phage suspensions were prepared as described by [2]. Three microliters of each sample was pipetted onto carbon coated 200 mesh copper grids and stained with 2% aqueous uranyl acetate. The samples were viewed using a LEO 912 Omega TEM at 120kV (Zeiss, Oberkochen, Germany) housed at the University of Cape Town Physics Department. Images were collected using a ProScan CCD camera.

Results and Discussion

Isolation, morphology and host range testing

Table 1 GVE3 host range

Bacterium	Strain	BGSC no.	Sensitivity to GVE3
Geobacillus stearothermophilus	ATCC 12980^T	9A20^T	-
Geobacillus thermoleovorans	DSM 5366T	96A1^T	-
Geobacillus thermoleovorans	DSM 7263	90A1	-
Geobacillus subterraneus	DSM 13552^T	91A1^T	-
Geobacillus subterraneus	SAM	91A2	-
Geobacillus thermodenitificans	DSM 465^T	94A1^T	-
Geobacillus thermoglucosidans	DSM 2542^T	95A1^T	+
Geobacillus toebii	DSM 14590^T	99A1^T	-
Geobacillus kaue	HU	105A1	-

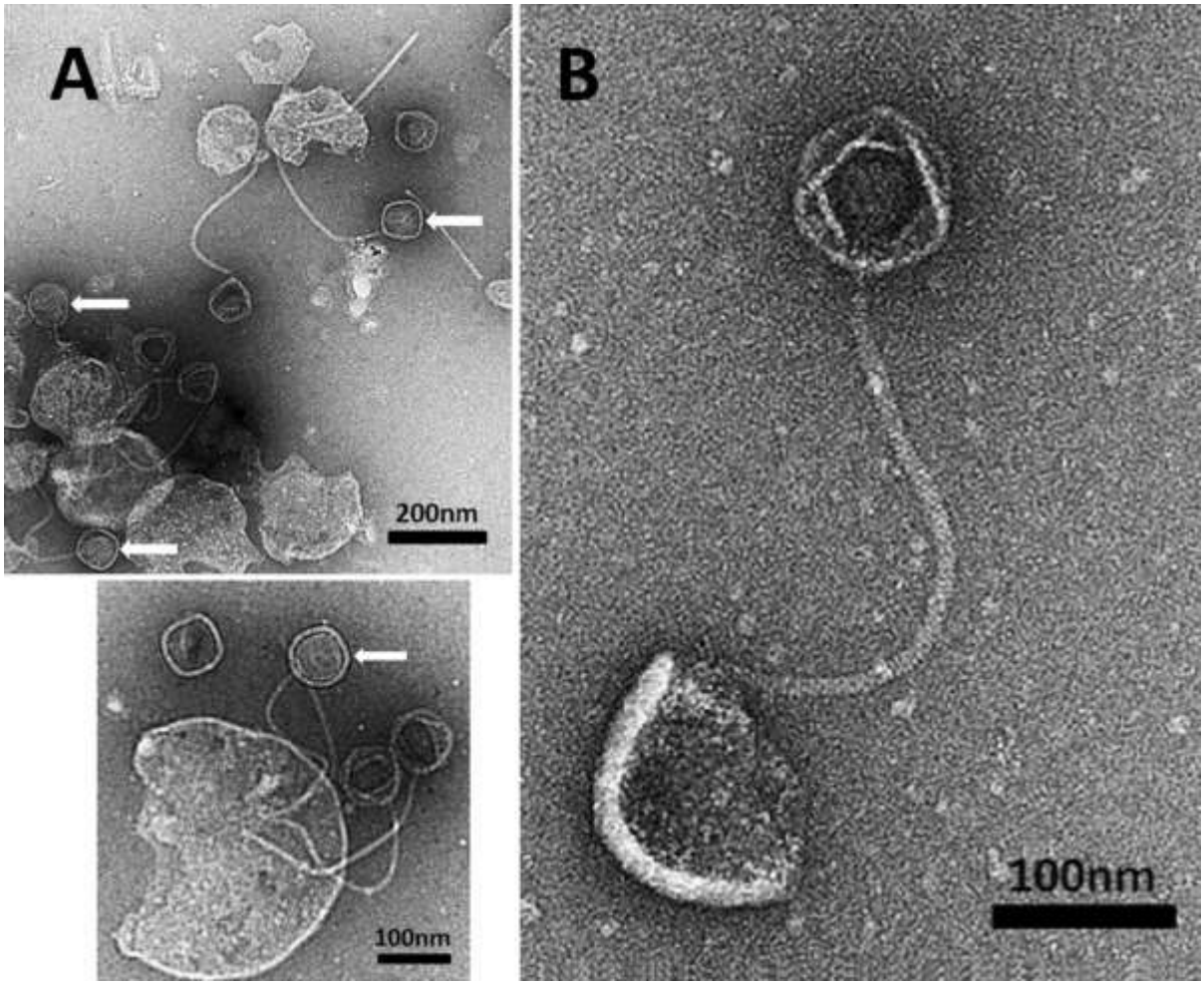

Fig. 1 Bright field TEM of phage GVE3. A) Lower- (top micrograph) and highermagnification (bottom micrograph) images of several phage attached to cell debris, including some that may still contain nucleic acid in the head (white arrows). B) High magnification image of a single phage particle

The phage was a donation from TMO Renewables. Transmission electron microscopy indicated that *G. thermoglucosidasius* phage GVE3 hadmorphological characteristics of the B1 morphotype group of the *Siphoviridae* family[1]with a non-contractile tail (± 210nm long) and isometric head (90nm – 100nm in diameter) (Fig. 1). GVE3 was tested for its ability to infect a range of *Geobacillus*species (Table 1), but was onlycapable of infected*G. thermoglucosidasius*.

The GVE3 genome

The GVE3 genome sequence was determined to be 141298bp in length and showed a much lower G+C content (29.6%) than its *G. thermoglucosidasius* host (44%), as is typical for most phage host pairs[64].It has been shown that higher AT content results in lower relative entropy (D_{KL}) of a DNA molecule which could be associated with structural changes in the molecule [12]. Perhaps the lower than average AT content of GVE3 plays a role in its adaptation to thermophily, or alternatively is a reflection of the energy cost of producing nucleotides for phage genome synthesis [64]. This genome size makes it the largest known*Geobacillus*-infecting phage. Overall, the GVE3 genome shares little nucleotide level identity with any bacteriophage genome currently on the NCBI database (as of 03-03-2015).However, small sections of the genome share significant nucleotide identity with other phage genomes (vB_BanS_Tsamsa, Spβc2, c-st) and *Geobacillus*, *Bacillus* and *Clostridium* genome sequences (Table S3).

A total of 202 putative open reading frames were identified, 62 of which could be assigned a function based on BLAST similarity to genes of known function. The GVE3 genome displays the classic modular arrangement seen in many other *Siphoviridae*(Fig. 2). G+C skew analysis indicated that a replication terminus could be located between the putative holin/endolysin (ORF53) genes and recombinase (ORF54) [65; c-st], while the origin of replication was predicted to lie at ±3700bp (Fig. 3).Repeat regions, often < 10 bp, are associated with regions where DNA replication is initiated,correspond to sites of gene regulation or transcription termination [10; 56; 61]. Depending on the search criteria, hundreds of inverted and direct repeats of < 10bp could be identified on the GVE3 genome, although their functional importance, if any, remains to be determined. A search for direct and inverted repeats of> 7bp and no more than 30bp apart with 100% nucleotide homology gave a total of 582 repeats. Two of these invert repeats (TATTTTTT / TAATTAT) are located immediately downstream of ORF3 and in the region predicted to be the origin of replication and may play a role in the initiation of replication.

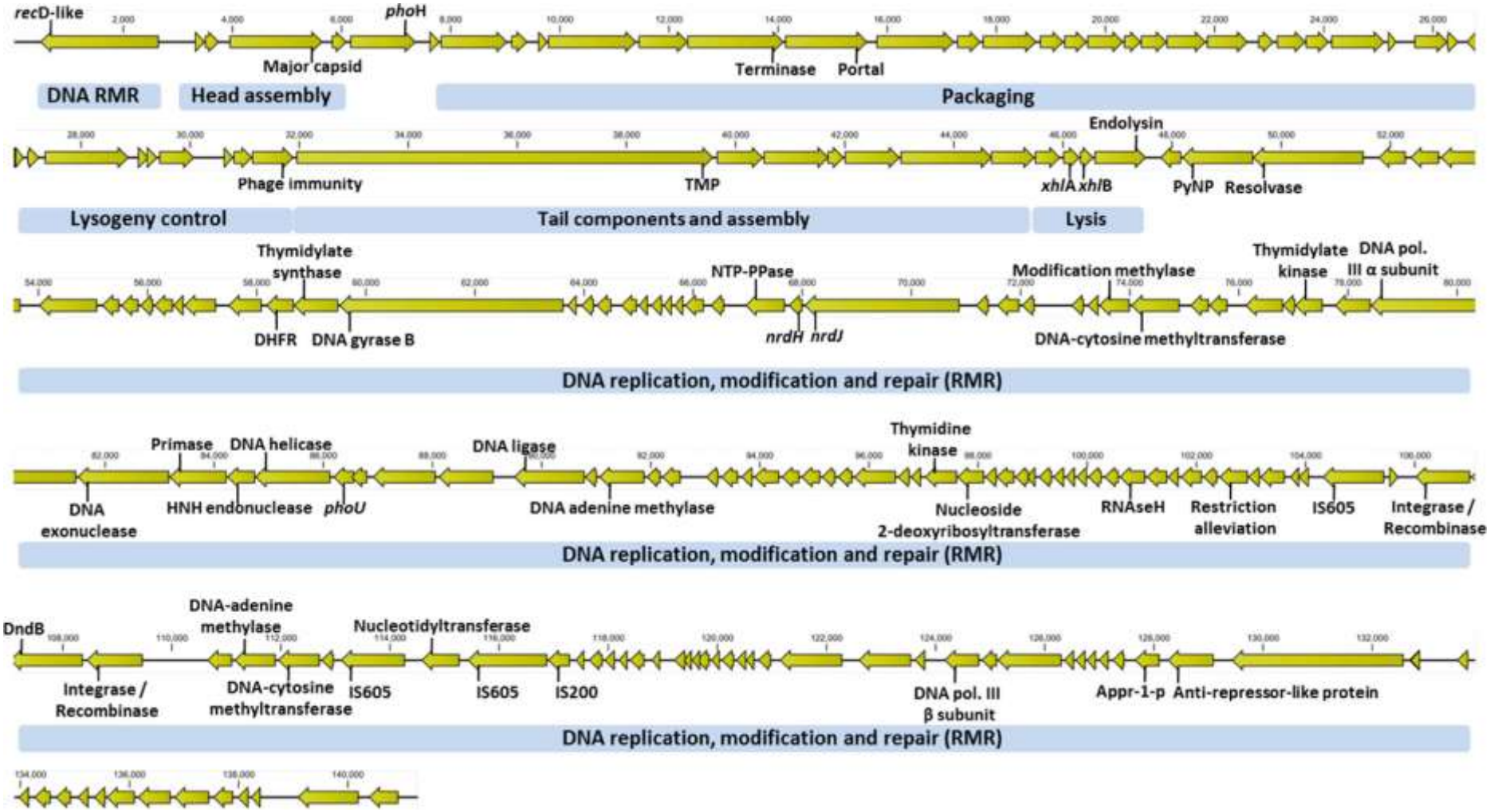

Fig. 2 GVE3 genomic arrangement. Blue boxes indicate modular areas

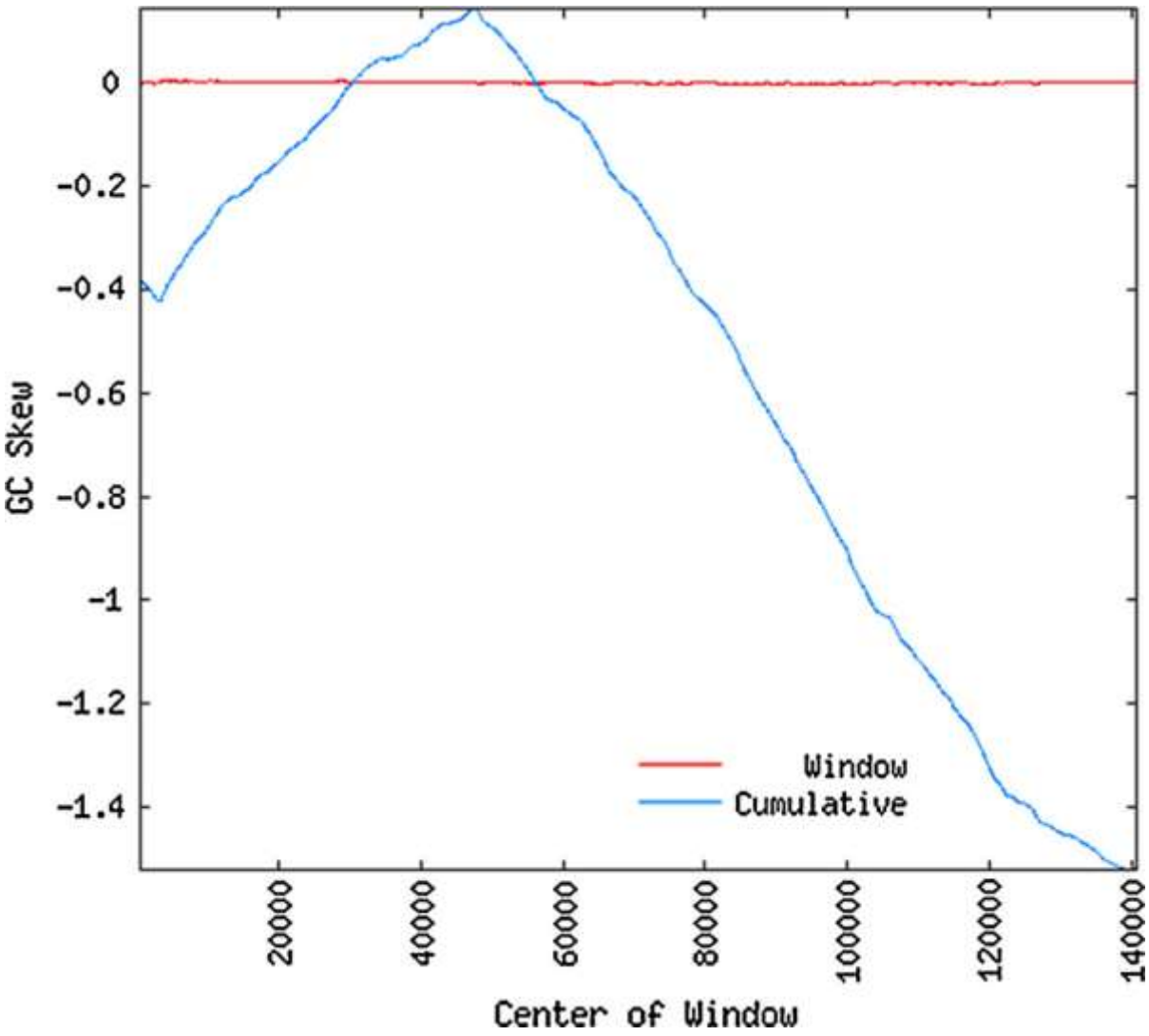

Fig. 3 GC skew analysis of the GVE3 genome showing putative replication origin (ori) and termination sites (ter) calculated using a window size of 1000 bp and a step size of 100 bp

Although GVE3 does not appear to encode any tRNA's, it does encode a putative ADP-ribose-1-monophosphatase (ORF184; Appr-1-p), an enzyme typically involved in tRNA splicingandencoded on a wide variety of phage genomes including vB_BanS-Tsamsa [25]. The exact role of this phage element is not clearly established[66], although the link with tRNA synthesis suggests that it could function to remove a rate limiting step in tRNA processing in the host or to aid in recycling of nucleotides [37].

The closest relatives to GVE3, based on sub-systems analysis using RAST, appear to be uncharacterized prophages from *Clostridiumthermocellum* and *Bacillus* species. The phages predicted to be the most closely related to GVE3, using PHAST (Table S4; http://tinyurl.com/mtg3fbs), are those from *Bacillus* (Spβc2; vB_BanS_Tsamsa) and *Clostridium*(c-st) rather than the known *Geobacillus*phages,an observation which is consistent with an analysis of the terminase large subunit (Fig.4). GVE3 thus appears to be most closely related to the recently described *B. anthracis* infecting vB_BanS_Tsamsa [25].

Tetranucleotide usage deviation (TUD) analysis gave a Pearson's correlation coefficient of 0.665 when comparing GVE3 to the genome of *G. thermoglucosidasius*. Interestingly, when comparing the GVE3 sequence to those of *Bacillus anthracis*and *Bacillus cereus,* significantly higher correlation coefficients were obtained (0.796 and 0.797, respectively). TUD analysis using all available *Geobacillus* species genome sequences (*G. kaustophilus, G. toebii, G. thermodinitrificans, G. themoleovorans, G. thermoglucosidasius, G. thermoglucosidans, G. stearothermophilus, G. subterraneus and G. caldoxylosilyticus*), demonstrated that GVE3's TUD was most closely matched to that of *G. toebii* (0.705).

Assuming that TUD analysis provides a measure of the adaptation of phage genomes to that of their hosts over time[60], the GVE3 TUD value does suggest that*G. thermoglucosidasius*may not be the prevalent host in nature.The higher correlation coefficientsof the GVE3 TUD when compared to*B. anthracis* and *B. cereus*(c.f., *G. thermoglucosidasius*) suggest that there may be an as yet unidentified *Geobacillus*specieswith TUD patterns more similar to these two *Bacillus* species that could be the "natural" hosts for GVE3. Alternatively, these results could suggest that GVE3 has "recently" evolved from a mesophilic counterpart and that the high TUD correlation to mesophilic*Bacillus* species is a genuine indication of its evolutionary heritage. A similar relationship has been observed for GBK2, a *G.kaustophilus*-infecting phage that is most closely related to the *Bacillus subtilis* phage SPP1[50].

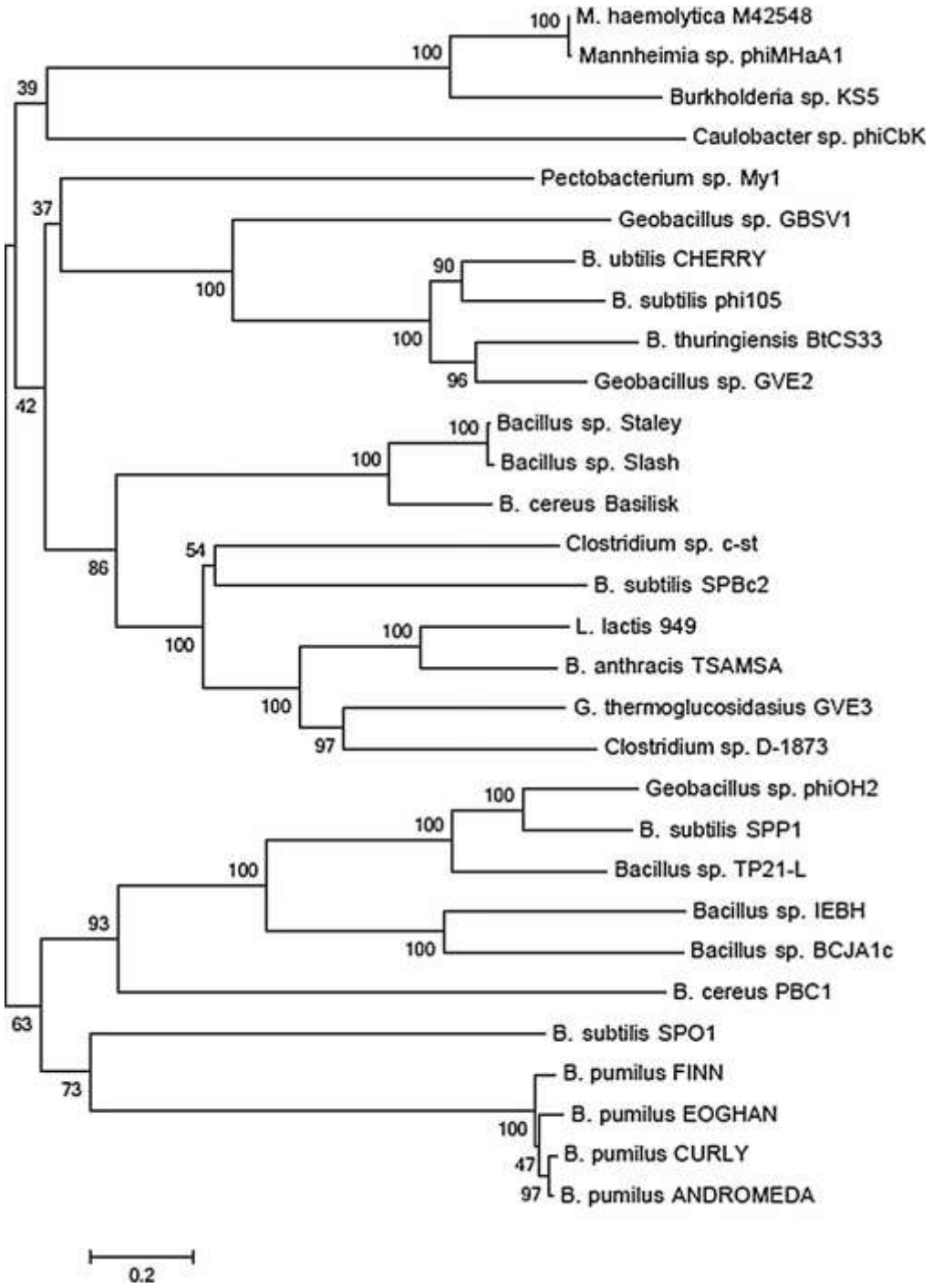

Fig. 4 Neighbor-joining tree comparing full length amino acid sequences of GVE3 terminase large subunit with related proteins. The optimal tree with the sum of branch length = 15.69592415 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and are in units of the number of amino acid substitutions per site (scale bar). The analysis involved 28 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 821 positions in the final dataset. GVE3, G. thermoglucosidasius (KP144388); IEBH, Bacillus sp. (NC_011167); BCJA1c, Bacillus sp. (NC_006557); TP21-L, Bacillus sp. (NC_011645); SPP1, B. subtilis (NC_004166); PBC1, B. cereus (NC_017976); phiOH2, Geobacillus sp. (NC_021784); D-1873, Clostridium sp. (ACSJ01000014); vB_BanS-Tsamsa, B. anthracis (NC_023007); 949, L. lactis (NC_015263); SPBc2, B. subtilis (AF020713); c-st. Clostridium sp. (D90210); Basilisk, B. cereus (KC595511); SPO1, B. subtilis (NC_011421); Slash, Bacillus sp. (NC_022774); Staley, Bacillus sp. (NC_022767); FINN, B. pumilus (NC_020480); EOGHAN, B. pumilus (NC_020477); ANDROMEDA, B. pumilus (NC_020478); CURLY, B. pumilus (NC_020479); BtCS33, B. thuringiensis (NC_018085); phi105, B. subtilis (NC_004167); CHERRY, B. anthracis (NC_007457); GBSV1, Geobacillus sp. (NC_008376); My1, Pectobacterium sp. (NC_018837); phiCbK, Caulobacter sp. (NC_019405); KS5, Burkholderia sp. (NC_015265); phiMHaA1, Mannheimia sp. (NC_008201); M2548, M. haemolytica (CP005383)

Evolution from mesophily to thermophilyshould involve the adaptation (in both thermophily and thermostability) of phage proteins, and it is therefore unlikely that the thermophilicGVE3 phage would be capable of replicating effectively in a mesophilic host.

DNA metabolism and replication

GVE3 encodes several proteins associated with nucleotide metabolism, including pyrimidine nucleoside phosphorylase (ORF55; PyNP), thymidylate synthase (ORF69; TS), thymidine kinase (ORF123; TK), ribonucleotidereductase (ORF82/83; RNR), nucleoside triphosphate pyrophosphohydrolase (ORF81; NTP-PPase) and nucleoside-deoxyribosyltransferase (ORF124; ND). Although the ORF encoding the putative RNR is most closely related to Class II RNR's, there is a small ORF directly downstream of this gene which shows high homology to a*nrdH*-like gene.Ribonucleotidereductases can be divided into several classes (Ia, Ib ,Ic; II and III) of which Class II RNR's are usually encoded by a single ORF (*nrdJ*), are oxygen independent and usually rely on vitamin B12 for generation of the tyrosyl radical *in vivo*[21]. Class Ib RNR's, encoded by *nrdHIEF*, rely on the glutaredoxin-like protein*nrdH*to generate the radical needed for catalysis [80]. GVE3 encodes a class IIribonucleotidereductase (*nrdJ*), as well as a component of a class Ib RNR, *nrdH*. The presence and spatial orientation of both *nrdJ*-like and *nrdH*-like ORF's would suggest that they function together. This unusual arrangement has been described for three *Mycobacterium*siphoviruses: Che12, D29 and L5 [21]. It has been argued that the *nrdH*homologue, in these genomes, was acquired through horizontal gene transfer. Phage genomes are, however, under strong selective pressure to remain within a strict size limit and all retainedgenes are expected to confer some metabolic advantage to the host and the phage[23]. In the case of GVE3, the proximity and spatial arrangement of *nrdH* and *nrdJ* as well as the retention of only the *nrdH* homolog (as opposed to any of the *nrdIEF* genes or gene fragments), would argue that these genes confer an advantage to the phage, perhaps *via* interaction with host encoded components.

The NTP-PPasecontains a MazG domain. MazG belongs to the family of allα-nucleoside triphosphate pyrophosphohydrolases, thought to be responsible for hydrolysis of all non-canonical nucleoside triphosphates produced as a by-product of metabolism and which are toxic to the host, into monophosphate derivatives, thus playing a house-cleaning role[13]. An alternative hypothesis is that, at least in*E. coli*, the NTP-PPasecontrols the levels of the global regulator ppGpp, redirecting transcription in favour of genes important for starvation

survival [47]. Homologues of these proteins have been identified on many phage genomes[28]. In *E. coli*,*mazG* is co-transcribed with a toxin-antitoxin system (*mazFE*) [27]. It is worth noting that GVE3 encodes several MazF/PemK homologues (ORF38, 40 and 185),although no *mazE* homologues could be identified, and the GVE3*mazG*-homologue is not co-transcribed with any of these. Whether or not the phage NTP-PPasefulfils multiple roles after host infection, such as regulating the levels ofMazF-like toxin produced or eliciting a host survival response to steer its metabolism towards viral production and / or removing toxic nucleoside triphosphates, remains to be determined.

Three DNA polymerase-like subunits are present on the GVE3 genome. Two of these (ORF97 and 176) are most closely related to the alpha- and beta-clamp subunits of the DNA polymerase III family, similar to those found on *Bacillus* phage vB_BanS_Tsamsa, *Clostridium*phages c-st, D-1873 and *Lactococcus* phage 949. The third subunit, ORF8, shows homology to DNA polymerase A. Other ORF's, the products of which may form part of the DNA Pol IIIholoenzyme,are a primase (ORF99) anda helicase (ORF101).It has been demonstrated for the *E. coli* DNA polymerase that only the alpha subunit is required for processive replication *in vitro*, although the authors conceded that other subunits,including subunit ε, may be required *in vivo* due to the polymerase encountering obstacles such as proteins bound to the DNA and DNA lesions not taken into account in their *in vitro* assay system [49]. As not all DNA polymerase III holoenzyme components could be identified on the GVE3 genome, it is possible that the phage recruits host-encoded subunits to complete the polymerase holoenzyme assembly to enable the highly processive DNA replication required for fast and accurate replication of the phage genome[14; 69].

3.4 Structural proteins

Aputative tail tape measure protein (TMP; ORF42/43) appears to be interrupted by a 310bp insertion (33537bp-33847bp), mostlikely a group I self-splicing intron as predicted by RNAweasel. As for phage JCL1032 from *Lactobacillus delbruckeii*[63], the 3' end of the ORF encoding the N-terminal protein sequence (31948bp-33536bp) ends with a TAG stop codon followed by the intron. Over the length of the putative TMP, seven large non-perfect amino acid repeats could be identified (≤102aa). The presence of repeat regions in these proteins has been reported previouslyand is thought to be of structural significance in determining tail length [8].

Mobile elements

Four putative integrase/recombinase genes were identified (ORF28, 54, 147 and 149), none ofwhich share significant amino acid similarity with each other, a feature noted with phage vB_BanS_Tsamsa[25]. The GVE3 phage genome carries three IS605 family OrfB genes (ORF145, 154 and 156). Insertion sequences of this family sometimes comprise two genes encoding an OrfA (IS200 family) and OrfB, together serving as the functional transposon [30]. One OrfB homologue on GVE3 (ORF156) does have an IS200 family gene directly upstream (ORF157), suggesting that they act co-ordinately. The arrangement of the genes is unusual in that they are transcribed in the same direction while most IS200 family transposons, when associated with an OrfB IS605 element, are divergently transcribed. Parts of GVE3 genome have been incorporated into *Geobacillustoebii* WCH70 CRISPR regions (Table S2). One of these spacers (36bp) is located in the sequence directly downstream of ORF143 on GVE3. Currently, the incorporation of sequences into CRISPR spacer regions is thought to occur through the identificationof bi- or trinucleotidesequences found adjacent to the protospacers which are eventually incorporated in the CRISPR array, andit is now thought that all Type I CRISPR systems target invading DNA for degradation [85]. Interestingly, an IS605/IS200 element (GWCH70_2010 and 2011) is situated directly upstream of the Cas6 (2068682bp-2069410bp) gene in WCH70, probably inactivating this CRISPR array. This CRISPR array also carries the 36bp spacer, and it is tempting to speculatethat a connection exists between these elements.The 36bp sequence may be important in the ability of the ORF143 transposon to jump, and incorporation of this spacer into a CRISPR cassette may inactivate the transposon, preventing it frominactivating host defence systems.

Nucleotide modifications

Digestion with several restriction endonucleases, including four base cutter *RsaI* for which there are 228 sites on the GVE3 genome, was not successful, whereas treatment with *Alu*I (335 sites) resulted in digestion of the DNA (Table S5). Examples where*Alu*Ibut not *Rsa*Iwould digest DNA have been reported, and are thought to be due tosubstitution of thymine with deoxyuridine or substitution of guanine with deoxyinosine[11]. It has also been established that *Alu*I cannot digest the following modified sites: ^{m6}AGCT, AG^{m4}CT, AG^{m5}CT,

AG^{hm5}CT[51],and these can probably be excluded as the modifications present in GVE3 DNA.The presence of putative methylases potentially targeting adenine and cytosine residues (ORF's 108, 151 and 152) as well as a DndB domain (ORF146) suggests that the phage DNA is modified to avoid digestion by host encoded enzymes. For example, *E. coli*T-even phage contains hydroxymethylcytosine (HMC) and *B. subtilis* phage PBS1 contains uracil in place of thymine. The pyrimidine, 5-hydroxymethyluracil (HMU), replaces thymine in *B. subtilis*phages SP8, SP5C, SPOl, SP82 and 4e [55]. GVE3 also encodes a putative restriction alleviation protein (102951bp-103163bp), possibly part of a strategy to avoid host defences.

The presence of restriction endonucleasesthat inhibit genetic transformation of *Geobacillus*species, and in particular *Hae*III in *G. thermoglucosidasius*has been reported (WO2006117536A1; Suzuki et al., 2012). Interestingly, all but one of the *Hae*III sites on GVE3, of which there are only 10, are located within the 3' terminal 946bp of the phage genome. They are irregularly spaced and do not appear to form part of conserved repeats. Digestion of phage DNA with *Hae*III could not be detected. The limited number of *Hae*III sites and their location may indicatethat the phage genome is under selective pressure to remove such sites. It is tempting to speculatethatas for phage P1, the 946bp region containing *Hae*III sites constitutes a *pac* site and that *pac*site cleavage is controlled by the methylation state surrounding the cleavage site [75].

GVE3 proteome

To confirm the expression of predicted ORF's, the complete proteome of GVE3 was determined. The protein products of all predicted ORF's listed in Table S1, except ORF5,60 and 169, could be identified by at least 3 unique peptides.The three segments of ORF60, which contains two frame shift mutations, are clearly similar to a hypothetical protein identified in a *Bacillus* species. (WP_028394443.1).However, no peptides similar to any of the three segments of the ORF could be identified and we conclude that ORF60 is an un-translated region.A peptide corresponding to the PyNPprotein was identified, suggestingthat thisenzyme may play a role in post-infection nucleotide metabolism (see below). No peptide sequences could be identified for the 310bp region predicted to be a group I self-splicing intron (33537bp-33847bp) indicating that this is likely to be an un-translated region. If the intron self-excises from this region once inside the host, it is reasonable to expect that a fusion protein, the functional TMP, would be formed by the N- and C-terminal regions of the predicted TMP interrupted by this intron. However, no evidence could be found for the formation of such a fusion protein between thesetwo terminal regions and it is likely that each ORF is expressed as a unique protein.The

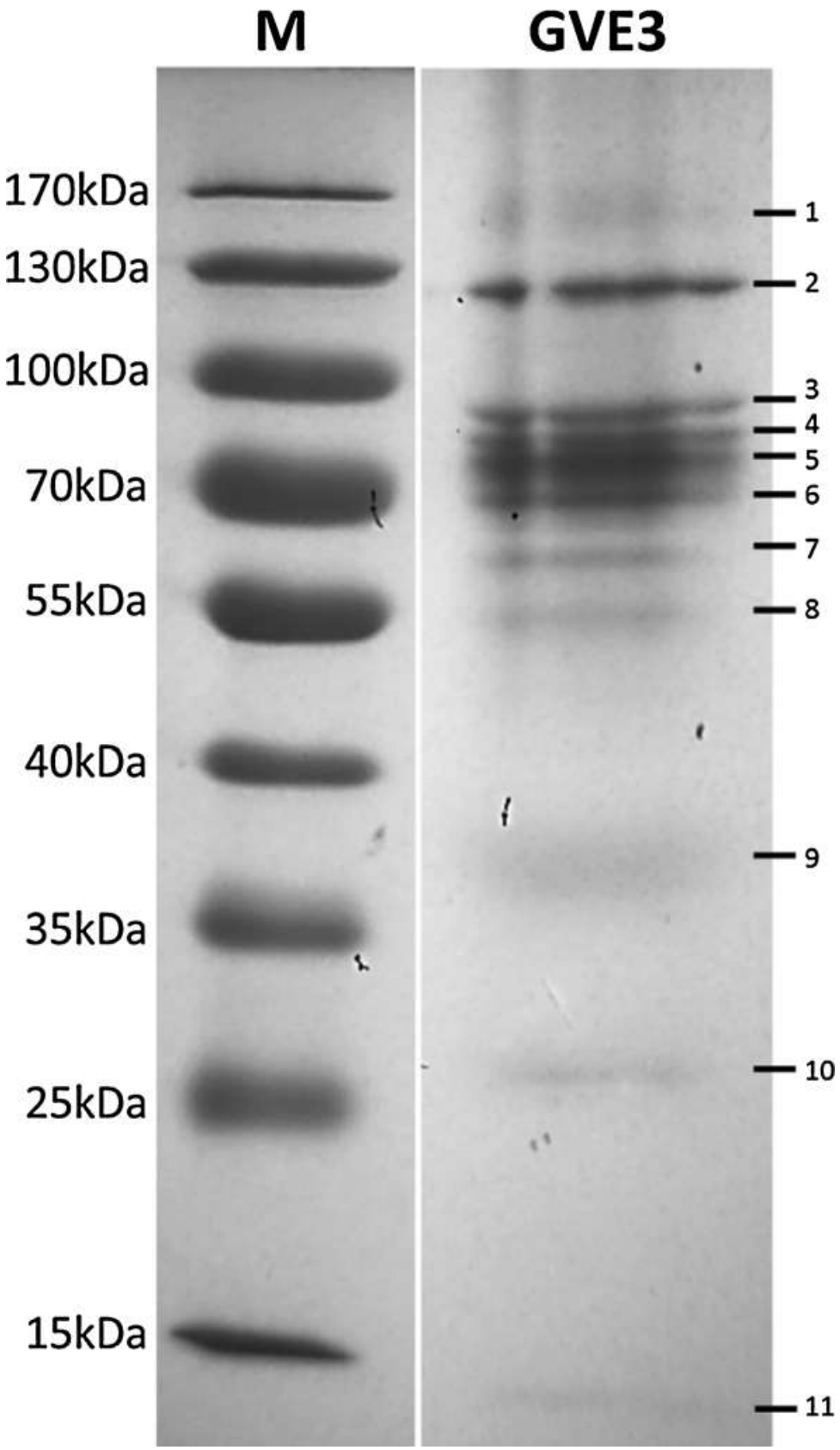

Fig. 5 SDS-PAGE of GVE3 structural proteins. M, Molecular mass marker

DNAsequence of ORF70contains a stop codon (TAG) in the reading frame which translates to VLD*EVK: the identification of a VLDEVK-containing peptide suggests either read-through translation or ribosome slippageoccurs over this codon. GVE3 structural proteins were also analysed by SDS-PAGE gel (Fig. 5). Eleven proteins could be identified of which band 6 corresponds to the size of the predicted major head protein (ORF4) while 7 and 8 are likely the N-terminal portion of the tape measure protein (ORF42) and the portal protein (ORF14) respectively.

Lysis andlysogeny

There are at least two holin homologues located directly upstream of the endolysin-encoding ORF, the second of these having what appears to be a dual start motif (M-X_n-M) with a lysine being one of the two residues separating the methionines. The arrangement of the genes and homology to *xhlA*/*xhlB*/*xlyA* genes from *B. subtilis* phage PBSX suggests that lysis might occur in a manner similar to that system [35]. ORF51 has one predicted transmembrane region (75-97aa) while ORF52 has two such regions (9-31aa; 41-59aa).

Initial plaque assays demonstrated “bulls-eye” plaque morphology, suggestive of host lysogeny[41]. Several bacterial colonies could be observed growing inside plaques. These were isolated and tested for their sensitivity to the phage and were found to be resistant to phage infection. The genome sequence for one of these isolates was determined (IlluminaMiSeq; 55 fold coverage) and served as confirmation of the GVE3 genome sequence obtained with Roche 454 sequencing. This showed that the phage genome had inserted into the bacterial host genome and that the *attB* site, with a 23bp sequence (GGTGGCGTCGGCGATACGACGAC) was duplicated on insertion (Fig.6). This sequence only occurs once in the *G. thermoglucosidasius* 11955 genome, located 247bp from the start of pyrimidine nucleoside phosphorylase (*deoA*), a region known to be interrupted by phage insertion in other genomes [16].The phage encodes a putative PyNP, downstream of a resolvase, in which the *attP* site is situated. Incorporation of GVE3 genome sequence in CRISPR spacer regions of the lysogen could not be detected, although two spacers with some nucleotide similarity to regions of the GVE3 genome were identified in the *G. thermoglucosidasius*11955 genome sequence (Table S2). Integration of the GVE3 genome sequence into that of its host is likely to inactivate the host-encoded PyNP. The presence of a phage-encoded PyNP could suggest an obligate requirement to retain this activity, and that once integrated, the host relies on the phage PyNP making use of a promoter located in the C-terminal region of the integrase (ORF55) or of read-through transcription from the promoter located upstream of the host encoded PyNP (Fig. 6). The PyNP on

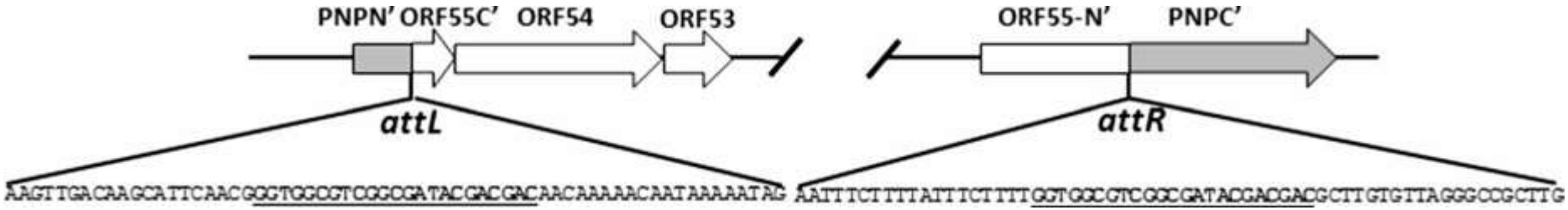

Fig. 6 Layout of the integrated phage. The space between the diagonal lines denotes the rest of the phage genome. The grey box and arrow represent the N- and C-termini, respectively, of the G. thermoglucosidasius pyrimidine nucleoside phosphorylase

GVE3 does not show 100% amino acid identity to the gene from *G. thermoglucosidasius*, or against any on the NCBInr database. If not essential for either phage or host (mutation in PyNP is non-lethal), it may suggest that GVE3 is a specialized transducing phage. No *G. thermoglucosidasius*genomic sequences were observed on the GVE3 genome or in the NGS data, suggesting that GVE3 is unlikely to be a generalized transducing phage. A putative anti-repressor protein (ORF183) , which contains an ORF6N domain and has amino acid similarity (50% over 112aa) with a truncated annotated anti-repressor protein in *Peptoclostridiumdifficile*, was identified [Table S1; 31]. In phage lambda this serves as part of the regulatory mechanism to switch between lysis and lysogeny. Early evidence, based on its overexpression in the host prior to infection, suggests that it plays the same role as in phage lambda (van Zyl et al., unpublished data).

Auxilliary metabolic genes

The GVE3 gene carries auxiliary metabolic gene *phoH*, and a putative regulator of *phoH* expression, *phoU*, is located upstream of the genes for DNA replication and distant (±79kb apart) to the *phoH* homologue. The phage also encodes a putative ADP-ribose-1-monophosphatase, which catalyses the conversion of ADP-ribose-1-monophosphate to ADP-ribose as part of the tRNA splicing pathway [37]. The role of *phoH* has not been clearly defined, with some studies demonstrating up-regulation under phosphate stress or phage infection [26]while others show down-regulation or no change. Should the GVE3 *phoH*gene expression be up-regulated, this might suggest that, as with other phages, DNA (and RNA) synthesis becomes rate limiting in the host once replication and transcription of the phage genome is initiated.

GVE3 signatures in *Geobacillus* genomes

Two regions of 100% nucleotide identity to CRISPR spacer regions were found in *G. toebii* WCH70 (Table S2). The presence of these nucleotide sequences suggests that GVE3 or a highly similar phage infected this strain in the past. PCR analysis using four primer pairs targeted to various areas of the GVE3 genome could not detect GVE3 in the chromosome of the *G. toebii*DSM 14590^{T} strain (Table 1), suggesting that superinfection immunity is unlikely to be the cause for failure to infect this strain. Several other putative GVE3-related sequences were identified in CRISPR repeats in a range of *G. thermoglucosidasius* genome sequences (Table S2).

Of the two spacers identified in the *G. toebii*WCH70 genome, one is located at the trailer end of the repeat region and the other, located in a second CRISPR array, at the leader end in that array, suggesting that this strain has been repeatedly challenged with the same phage [29; 84]. The absence of evidence of lysogenic integration of the GVE3 genome in the WCH70 genome could be due to CRISPR-mediated killing of the hosts contain an integrated phage or that have been infected in the past [22; 48]. Imperfect match spacers similar to GVE3 in CRISPR arrays on the 11955 genome could suggest infection by a closely related phage as seen in polyclonal phage populations during phage blooms or adaptation by the phage to circumvent CRISPR resistance [48; 84]. We suggest thatGVE3 represents the latest iteration of a much older version of the phage not currently targeted by the CRISPR system in *G. thermoglucosidacius*11955. Insertion of spacer sequences based on those identified in WCH70 could be used to engineer resistance by incorporating these into one of the 11955 CRISPR arrays [52].

Conclusion

GVE3, although a member of the well-known Siphovirus family and unremarkable with respect to the overall layout of genes and the genes encoded, appears to have a unique genome sequence with no closely related members in the current databases. Although there are indications that it may have had the capacity to infect other *Geobacillus* species in the past, the current specificity appears to be restricted to *G. thermoglucosidasius*.The relationships between the GVE3 genome and those of mesophilicphages and bacteria may be a consequence of the small number of thermophilic phage genome sequences in the databases, but may reflect the evolutionary history of a phage in transition from mesophily to thermophily. GVE3 encodes anumber of enzymes, including ATP dependent DNA ligase, DNA polymerase III, RNaseH, PyNP, holin and endolysin[78],all of which should be thermostable. These could be of commercial value or employed as research tools,such as in the use of endolysinin the treatment of milk to kill *Geobacillus* species spoilage organisms [15; 81]. *G. thermoglucosidasius* has been engineered as a platform organism for ethanol production and other industrial products but to date there is no mechanism for the introduction of large DNA fragments (>12kb) and GVE3 could potentiallybe developed as a system for introduction of novel or engineered metabolic and biosynthetic pathways.

Acknowledgements

The authors wish to thank TMO Renewables for the gift of the GVE3 phage. This work was funded by the National Research Foundation (NRF) of South Africa. The authors declare no conflict of interest.

References

1. Ackermann HW(2007) 5500 Phages examined in the electron microscope. Arch. Virol.152:227-243
2. Ackermann HW,Heldal M(2010) Chapter 18: Basic electron microscopy of aquatic viruses In Manual Of Aquatic Viral Ecology, American Society of Limnology and Oceanography, Inc., p 182-192
3. Adriaenssens EM, van Zyl LJ, de Maayer P, Rubagotti E, Rybicki E, Tuffin M, Cowan DA (2014)Metagenomic analysis of the viral community in Namib Desert hypoliths. Environ. Microbiol. doi:10.1111/1462-2920.12528
4. Ahn D-G, Kim S-I, Rhee J-K, Kim KP, Pan J-G, Oh J-W (2006) TTSV1, a new virus-like particle isolated from the hyperthermophiliccrenarchaeote*Thermoproteustenax*. Virol. 351:280-290
5. Altschul SF, Gish W, Miller W, Myers EW,Lipman DJ(1990)Basic local alignment search tool. J. Mol. Biol.215:403-410
6. Arnold HP, Zillig W, Ziese U, Holz I, Crosby M, Utterback T, Weidmann JF, Kristjanson JK, Klenk HP, Nelson KE, Fraser CM (2000) A Novel Lipothrixvirus, SIFV, of the Extremely ThermophilicCrenarchaeon*Sulfolobus*. Virol. 267:252-266
7. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics 9:75
8. Belcaid M, Bergeron A, Poisson G(2011)The evolution of the tape measure protein: units, duplications and losses. BMC Bioinform. 12:S10
9. Betley JN, Frith MC, Graber JH, Choo S, Deshler JO (2002) A ubiquitous and conserved signal for RNA localization in chordates. Curr. Biol. 12:1756-1761

10. BlatnyJM,GodagerL,Lunde M,Nes IF(2004)Complete genome sequence of the *Lactococcuslactis* temperate phage φLC3: comparative analysis of φLC3 and its relatives in lactococci and streptococci. Virology 318:231-244
11. Bodnarz JW,Zempsky W, Warder D, Bergson C, Ward DC(1983) Effect of nucleotide analogs on the cleavage of DNA by the restriction enzymes *Alu*I, *Dde*I, *Hin*fI, *Rsa*I, and *Taq*I. J. Biol. Chem. 258:15206-15213
12. Bohlin J, van Passel MWJ,Snipen L,Kristoffersen AB,Ussery D, Hardy SP(2012) Relative entropy differences in bacterial chromosomes, plasmids, phages and genomic islands. BMC Genomics13:66-78
13. Bryan MJ, Burroughs NJ, Spence EM, Clokie MRJ, Mann NH, BryanSJ (2008) Evidence for the intense exchange of MazG in marine cyanophages by horizontal gene transfer. PLOS One 3:doi:10.1371/journal.pone.0002048
14. Bullard JM, Williams JC, Acker WK, Jacobi C, Janjic N, McHenry CS (2002) DNA polymerase III holoenzyme from *Thermus thermophiles* identification, expression, purification of components, and use to reconstitute a processivereplicase. J. Biol. Chem. 277:13401-13408
15. Burgess SA, Lindsay D, Flint SH(2010)Thermophilic bacilli and their importance in dairy processing. Int. J. Food Microbiol. 144:215-225
16. Buxton RS, Hammer-Jespersen K, Hansen TD (1978) Insertion of bacteriophage lambda into the *deo* operon of *Escherichia coli* K-12 and isolation of plaque-forming λ*deo*$^{+}$ transducing bacteriophages. J. Bacteriol. 136:668-681
17. Chen Y, Wei D, Wang Y, Zhang X (2013) The role of interactions between bacterial chaperone, aspartate aminotransferase, and viral protein during virus infection in high temperature environment: the interactions between bacterium and virus proteins. BMC Microbiol. 13:48
18. Clokie MRJ, Millard AD,Letarov AV,Heaphy S (2011) Phages in nature. Bacteriophage1:31-45
19. Cripps RE,Eley K, Leak DJ, Rudd B, Taylor M, Todd M,Boakes S, Martin S, Atkinson T(2009) Metabolic engineering of *Geobacillusthermoglucosidasius* for high yield ethanol production. Metab. Eng.11:398-408
20. Doi K, Mori K, Martono H, Nagayoshi Y, Fujino Y, Tashiro K, Kuhara S, Ohshimaa T (2013)Draft Genome Sequence of *Geobacilluskaustophilus*GBlys, a Lysogenic Strain with Bacteriophage OH2. Genome Announc.1:e00634-13
21. Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M (2013) A bioinformatic analysis of ribonucleotidereductase genes in phage genomes and metagenomes. BMC Evol. Biol. 13:33

22. Edgar R, Qimron U (2010) The *Escherichia coli*CRISPR system protects from lysogenization, lysogens, and prophage induction. J. Bacteriol. 192:6291-6294
23. Feiss M,Siegele DA (1979) Packaging of the bacteriophage lambda chromosome: Dependence of *cos* cleavage on chromosome length. Virology 92:190-200
24. Felsenstein J (1985) Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791
25. Ganz HH, Law C, Schmuki M, Eichenseher F, Calendar R, Loessner MJ, Getz WM, Korlach J, Beyer W, Klumpp J (2014) Novel giant Siphovirus from *Bacillus anthracis*features unusual genome characteristics. PLoS One. 9:e85972
26. Goldsmith DB,Crosti G,Dwivedi B, McDaniel LD,Varsani A,Suttle CA,Weinbauer MG,Sandaa RA,Breitbart M(2011) Development of phoH as a novel signature gene for assessing marine phage diversity. Appl. Environ. Microbiol.77:7730-7739
27. Gross M, Marianovsky I, Glaser G (2006)MazG – a regulator of programmed cell death in *Escherichia coli*. Mol. Microbiol. 59:590-601
28. Hargreaves KR, Kropinski AM, Clokie MRJ (2014) Bacteriophage behavioral ecology How phages alter their bacterial host's habits. Bacteriophage 4:doi: 10.4161/bact.29866
29. Heler R, Marraffini LA,Bikard D(2014) Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. Mol. Microbiol. 93:doi:10.1111/mmi.12640
30. Höök-Nikanne J, Berg DE, Peek Jr. RM, Kersulyte D, Tummuru MKR, Blaser MJ (1999) DNA sequence conservation and diversity in transposable element IS605 of *Helicobacter pylori*. Helicobacter 3:79-85
31. Iyer LM, Koonin EV, Aravind L (2002) Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. Gen. Biol. 3:research0012.1–0012.11
32. Jin M, Ye T, Zhang X (2013) Roles of bacteriophage GVE2 endolysin in host lysis at high temperatures. Microbiol. 159:1597-1605
33. Jin M, Chen Y, Xu C, Zhang X (2014) The effect of inhibition of host MreB on the infection of thermophilic phage GVE2 in high temperature environment. Sci. Rep. 4:4823
34. Kotze AA,Tuffin IM, Deane SM, Rawlings DE(2006) Cloning and characterization of the chromosomal arsenic resistance genes from *Acidithiobacilluscaldus* and enhanced arsenic resistance on conjugal transfer of *ars* genes located on transposon TnAtcArs. Microbiology 152:3551-3560

35. Krogh S, Jørgensen ST, Devine KM(1998)Lysis genes of the *Bacillus subtilis* defective prophage PBSX. J. Bacteriol. 180:2110–2117
36. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. 305:567-580
37. Kumaran D,Eswaramoorthy S, Studier FW,Swaminathan S(2005) Structure and mechanism of ADP-ribose-1-monophosphatase (Appr-1-pase), a ubiquitous cellular processing enzyme. Prot. Science14:719-726
38. Lang BF, Laforest MJ, Burger G (2007)Mitochondrial introns: a critical view. Trends Genet. 23:119-125
39. Laslett D,Canback B(2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucl. Acids Res.32:11-16
40. Le Romancer M, Gaillard M, Geslin C, Prieur D (2007) Viruses in extreme environments. p99-113. In Life in Extreme Environments, Eds. Ricardo Amils, Cynan Ellis-Evans, Helmut Hinghofer-Szalkay. Springer, Netherlands
41. Levine M, Truesdall S, Ramakrishan T, Bronson MJ (1975) Dual control of lysogeny by bacteriophage P22: An antirepressor locus and its controlling elements. J. Mol. Biol. 91:421-438
42. Lin PP, Rabe KS, Takasumi JL, Kadisch M, Arnold FH, Liao JC (2014)Isobutanol production at elevated temperatures in thermophilic*Geobacillusthermoglucosidasius*. Metab. Eng. 24:1-8
43. Liu B, Wu S, Song Q, Zhang X, Xie L (2006) Two Novel Bacteriophages of Thermophilic Bacteria Isolated from Deep-Sea Hydrothermal Fields. Curr.Microbiol. 53:163-166
44. Liu B, Zhang X, (2008) Deep-sea thermophilic*Geobacillus* bacteriophage GVE2 transcriptional profile and proteomic characterization of virions. Appl. Microbiol. Biotechnol.80:697–707
45. Liu B, Zhou F, Wu S,Xu Y, Zhang X (2009)Genomic and proteomic characterization of a thermophilic*Geobacillus* bacteriophage GBSV1. Res. Microbiol. 160:166-170
46. Lowe TM, Eddy SR (1997)tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucl. Acids Res. 25:955-964
47. Magnusson LU, Farewell A, Nystrom T (2005)ppGpp: a global regulator in *Escherichia coli*. Trends in Microbiol. 13:236–242
48. Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. 11:181–190
49. Marians KJ, Hiasa H, Kim DR, McHenry C (1998) Role of the core DNA polymerase III subunits at the replication fork: Alpha is the only subunit required for processive replication. J. Biol. Chem. 273:2452-2457

50. Marks TJ, Hamilton PT (2014) Characterization of a thermophilic bacteriophage of *Geobacilluskaustophilus*. Arch. Virol. 159:2771-2775

51. McClelland M, Nelson M,Raschke E (1994) Effect of site-specific modification on restriction endonucleases and DNA modification methyltransferases. Nucl. Acids Res. 22:3640-3659

52. Millen AM, Horvath P, Boyaval P, Romero DA (2012) Mobile CRISPR/Cas-Mediated bacteriophage resistance in *Lactococcuslactis*. PLoS One. 7:e51663

53. Moser MJ, DiFrancesco RA, Gowda K, Klingele AJ, Sugar DR, Stocki S, Mead DA, Schoenfeld TW (2012)Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme. PLOS One. 7:doi:10.1371/journal.pone.0038371

54. Nazina TN, Tourova TP, Poltaraus AB, Novikova EV, GrigoryanAA, Ivanova AE, Lysenko AM,Petrunyaka VV, Osipov GA, Belyaev SS, Ivanov MV (2001) Taxonomic study of aerobic thermophilic bacilli: descriptions of *Geobacillussubterraneus* gen. nov., sp. nov.and*Geobacillusuzenensis* sp. nov. from petroleum reservoirs and transfer of *Bacillus stearothermophilus*, *Bacillus thermocatenulatus*, *Bacillus thermoleovorans*, *Bacillus kaustophilus*, *Bacillus thermoglucosidasius* and *Bacillus thermodenitrificans* to *Geobacillus*as the new combinations *G. stearothermophilus*, *G. thermocatenulatus*, *G. thermoleovorans*, *G. kaustophilus*, *G. thermoglucosidasius* and *G. thermodenitrificans*. IJSEM 51:433–446

55. Neubort S, Marmur J (1973) Synthesis of the unusual DNA of *Bacillus subtilis* bacteriophage SP-15. J. Virol. 12:1078-1084

56. Østergaard S, Brøndsted L, Vogensen FK (2001) Identification of a replication protein and repeats essential for DNA replication of the temperate lactococcal bacteriophage TP901-1. Appl. Environ. Microbiol. 67:774-781

57. Payeta JP,Suttle CA (2013)To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. Limnol.Oceanogr. 58:465–474

58. Peduzzi P, Gruber M, Gruber M, Schager M (2014)The virus's tooth: cyanophages affect an African flamingo population in a bottom-up cascade. ISME J. 8:1346-1351

59. Plotka M, Kaczorowska A-K, Stefanska A, Morzywolek A, Fridjonsson OH, Dunin-Horkawicz S, Kozlowski L, Hreggvidsson GO, Kristjansson JK, Dabrowski S, Bujnicki JM,Kaczorowskia T (2013) Novel highly thermostableendolysin from *Thermusscotoductus* MAT2119 bacteriophage Ph2119 with amino acid sequence similarity to Eukaryotic peptidoglycan recognition proteins. Appl. Environ. Microbiol. 80:886-895

60. Pride DT, Wassenaar TM, Ghose C, Blaser MJ (2006) Evidence of host-virus co-evolution in tetranucleotide usage patterns of bacteriophages and eukaryotic viruses. BMC Genomics 7:doi:10.1186/1471-2164-7-8
61. Quiles-Puchalt N, Tormo-Más MA, Campoy S, Toledo-Arana A, Monedero V, Lasa I, Novick RP, Christie GE, Penadés JR (2013) A super-family of transcriptional activators regulates bacteriophage packaging and lysis in Gram-positive bacteria. Nucl. Acids Res. 41:7260-7275
62. Rice G, Stedman K, Snyder J, Wiedenheft B, Willits D, Brumfield S, McDermott T, Young MJ (2001) Viruses from extreme thermal environments. Proc. Nat. Acad. Sci. 98:13341-13345
63. Riipinen KA, Alatossava T (2004) Two self-splicing group I introns interrupt two late transcribed genes of prolate-headed *Lactobacillus delbrueckii* phage JCL1032. Arch. Virol. 149:2013-2024
64. Rocha EPC, Danchin A (2002) Base composition bias might result from competition for metabolic resources. TRENDS in Genet. 18:291-294
65. Sakaguchi Y, Hayashi T, Kurokawa K, Nakayama K, Oshima K, Fujinaga Y, Ohnishi M, Ohtsubo E, Hattori M, Oguma K (2005) The genome sequence of *Clostridium botulinum* type C neurotoxin-converting phage and the molecular mechanisms of unstable lysogeny. Proc. Nat. Acad. Sci. 102:17472-17477
66. Savalia D,Westblade LF,Goel M,Florens L, Kemp P,Akulenko N,Pavlova O,Padovan JC,Chait BT, Washburn MP, Ackermann HW,Mushegian A,Gabisonia T,Molineux I,Severinov K(2008)Genomic and proteomic analysis of phiEco32, a novel *Escherichia coli*phage. J. Mol. Biol. 377:774-789
67. Schmidt TR, Scott II EJ, Dyer DW(2011) Whole-genome phylogenies of the family Bacillaceae and expansion of the sigma factor gene family in the *Bacillus cereus* species-group. BMC Genomics 12:430
68. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D (2008) Assembly of viral metagenomes from yellowstonehot springs. Appl. Environ. Microbiol. 74:4164-4174
69. Seco E,Zinder J,Manhart CM, Piano AL, McHenry C,Ayora S(2013) Bacteriophage SPP1 in vitro DNA replication strategies promote viral and disable host replication. Nucl.Acid. Res. 41:1711-1721
70. Sevostyanova A, Djordjevic M, Kuznedelov K, Naryshkina T, Gelfand MS, Severinov K, Minakhin L (2007) Temporal regulation of viral transcription during development of *Thermusthermophilus* bacteriophage φYS40. J. Mol. Biol. 366:420–435
71. Sime-Ngando ST, Lucas S, Robin A, Tucker KP, Colombet J, Bettarel Y, Desmond E, Gribaldo S, Forterre P, Breitbart M, Prangishvili D (2010) Diversity of virus–host systems in hypersaline Lake Retba. Environ. Microbiol. 8:1956-1972

72. Song Q, Zhang X (2008) Characterization of a novel non-specific nuclease from thermophilic bacteriophage GBSV1. BMC Biotechnol. 8:43

73. Song Q, Ye T, Zhang X (2011) Proteins responsible for lysogeny of deep-sea thermophilic bacteriophage GVE2 at high temperature. Gene 479:1-9

74. Sorokin DY, Berben T, Melton ED, Overmars L, Vavourakis CD, Muyzer G (2014) Microbial diversity and biogeochemical cycling in soda lakes. Extremophiles 18:791-809

75. Sternberg N, Coulby J (1990) Cleavage of the bacteriophage P1 packaging site (*pac*) is regulated by adenine methylation. Proc. Natl. Acad. Sci. 87:8070-8074

76. Suttle CA (2005) Viruses in the sea. Nature 437:356–361

77. Suzukim H, Yoshida K (2012) Genetic transformation of *Geobacilluskaustophilus* HTA426 by conjugative transfer of host-mimicking plasmids. J. Microbiol. Biotechnol. 22:1279-1287

78. Szekera K, Zhou X, Schwab T, Casanueva A, Cowan D, Mikhailopulo IA, Neubauer P (2012)Comparative investigations on thermostable pyrimidine nucleoside phosphorylases from *Geobacillusthermoglucosidasius* and *Thermus thermophiles*. J. Mol. Cat B: Enzymatic 84:27-34

79. Taylor MP, Eley KL, Martin S, Tuffin MI, Burton SG, Cowan DA (2009)Thermophilicethanologenesis: future prospects for second-generation bioethanol production. Trends Biotechnol. 27:398-405

80. Torrents E (2014)Ribonucleotidereductases: essential enzymes for bacterial life. Front. Cell. Infect. Microbiol. 4:doi: 10.3389/fcimb.2014.00052

81. Viedma PM, Abriouel H, Omar NB, Lopez RL, Valdivia E, Gálvez A (2009) Inactivation of *Geobacillusstearothermophilus* in canned food and coconut milk samples by addition of enterocin AS-48. Food Microbiol. 26:289-293

82. Wang Y, Zhang, X (2008) Identification and characterization of a novel thymidylate synthase from deep-sea thermophilic bacteriophage *Geobacillus* virus E2. Virus Genes 37:218–224

83. Wang Y, Zhang X (2010) Genome Analysis of Deep-Sea Thermophilic Phage D6E. Appl. Environ. Microbiol. 76:7861-7866

84. Weinberger AD, Sun CL, PlucinskiMM, Denef VJ, Thomas BC, Horvath P, Barrangou R, Gilmore MS, Getz WM, Banfield JF (2012) Persisting viral sequences shape microbial CRISPR based immunity. PLoS One. 8:e1002475

85. Westra ER, Swarts DC, Staals RHJ, Jore MM, Brouns SJJ, van der Oost J (2012) The CRISPRs, they are A-Changin': How prokaryotes generate adaptive immunity. Annual Rev. Genet. 46:311-339

86. Zhou Y, Liang Y, Lynch K, Dennis JJ,Wishart DS (2011) PHAST: A Fast Phage Search Tool. Nucl. Acids Res. 39:347-352

87. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. Edited in Evolving Genes and Proteins by V. Bryson and H.J. Vogel, pp. 97-166. Academic Press, New York.

Supplementary material

Table S1. Predicted open reading frames on GVE3 and closest BLASTp hit on the NCBI database

ORF number	Size in amino acids	Start and end positions bp	Selected BLAST hits and comments; accession number; (length of protein on database in aa)	% Identity/Similarity (over number of aa)
1	724	-(478-2649)	RecD/TraA family helicase *Bacillus cereus* WP_016094953.1 (742)	46/66 (338/486)
2	61	+(3313-3495)	Hypothetical protein; phage SPbeta *Bacillus amyloliquefaciens* subsp. plantarum UCMB5033 YP_008413102.1 (59) / SPBc2 prophage-derived protein YonP *Bacillus amyloliquefaciens* LL3 YP_005546102.1 (59)	41/73 (19/34) 42/75 (19/34)
3	87	+(3495-3755)	Hypothetical protein *Bacillus mycoides* WP_003204337.1 (92)	56/73 (48/63)
4	557	+(3947-5647)	Conserved hypothetical *Salsuginibacillus kocurii* WP_018923841.1 (556) / Major head protein and HOOK domain	34/55 (186/300)
5	91	+(5821-6093)	DNA-binding protein HU-alpha *Anoxybacillus flavithermus* WK1 YP_002315451.1 (101)	74/82 (67/74)
6	400	+(6158-7357)	Hypothetical protein *Ruminococcus gnavus* WP_004840076.1 (415) / PhoH family protein *Spirochaeta africana* DSM 8902 YP_005476454.1 (434)	35/56 (144/235) 29/50 (124/217)
7	64	+(7612-7803)	Hypothetical protein *Bacillus licheniformis* WP_021837902.1 (67) / SPBc2 prophage-derived protein YonK *Bacillus amyloliquefaciens* LL3 YP_005546104.1 (63)	57/79 (36/50) 56/77 (34/47)
8	407	+(7816-9036)	Hypothetical protein *Bacillus licheniformis* WP_021837901.1 (405) / DNA polymerase I *Bacillus* phage Troll YP_008430961.1 (431)	38/58 (152/233) 25/49 (63/126)
9	98	+(9110-9406)	Excinuclease ABC subunit A Candidatus *Protochlamydia amoebophila* WP_011176187.1 (1900)	30/52 (22/38)
10	59	+(9604-9780)	Hypothetical protein *Coprococcus sp*. HPP0074 WP_016438666.1 (57)	41/70 (22/38)
11	539	+(9785-11401)	SPBc2 prophage-derived protein YomD *Bacillus megaterium* WSH-002 YP_005494512.1 (486)	28/45 (116/189)
12	301	+(11444-12346)	Hypothetical protein *Bacillus cereus* WP_016094960.1 / SPBc2 prophage-derived protein YonG *Bacillus sonorensis* WP_006640189.1	37/56 (97/149) 28/46 (70/115)
13	589	+(12336-14102)	Hypothetical protein Bacillus cereus WP_016094961.1 (585) / Putative terminase ATPase subunit *Lactococcus* phage 949 YP_004306307.1 (565)	43/63 (245/362) 37/55 (204/308)
14	502	+(14133-15638)	Hypothetical protein SPBc2p055 *Bacillus* phage SPBc2 NP_046607.1 (506) / C-terminal portal protein domain HK97	26/46 (126/230)
15	475	+(15803-17227)	Hypothetical protein CA_C1132 *Clostridium acetobutylicum* ATCC 824 NP_347765.1 (488) / C-terminal Smc domain; cell division and chromosome partitioning protein	27/49 (123/224)
16	145	+(17280-17714)	Hypothetical protein CA_C1131 *Clostridium acetobutylicum* ATCC 824 NP_347764.1 (147)	36/53 (51/76)
17	327	+(17751-18731)	Hypothetical protein CA_C1130 *Clostridium acetobutylicum* ATCC 824 NP_347763.1 (339)	37/58 (124/196)
18	143	+(18801-19229)	Hypothetical protein *Paenibacillus* WP_009671518.1 (165)	27/54 (32/65)
19	138	+(19244-19657)	Hypothetical protein *Blautia hansenii* WP_003020132.1 (142)	44/62 (52/74)
20	220	+(19667-20326)	Hypothetical protein BATR1942_07955 *Bacillus atrophaeus* 1942 YP_003973461.1 (223)	24/48 (52/106)
21	104	+(20341-20652)	Hypothetical protein SERP1632 *Staphylococcus epidermidis* RP62A YP_189197.1 (105)	35/59 (33/55)

22	158	+(20649-21122)	Hypothetical protein *Lysinibacillus sphaericus* WP_010858764.1 (180)	27/46 (46/81)
23	239	+(21122-21838)	Hypothetical protein *Clostridium* phage D-1873 WP_003377629.1 (247)	35/59 (67/113)
24	250	+(21857-22606)	Hypothetical protein *Paenibacillus dendritiformis* WP_006675026.1 (256)	39/59 (98/157)
25	96	+(22796-23083)	Hypothetical protein *Streptococcus suis* WP_024395613.1 (106)	33/51 (29/45)
26	175	+(23143-23667)	SPBc2 prophage-derived protein YomO *Bacillus amyloliquefaciens* LL3 YP_005545386.1 (162) / C-terminal AAK aspartokinase-like domain	24/53 (38/84)
27	142	+(23668-24093)	SPBc2 prophage-derived protein YomN *Bacillus amyloliquefaciens* subsp. plantarum YAU B9601-Y2 YP_005421066.1 (138) / N-terminal Clusterin domain	46/62 (63/86)
28	326	+(24138-25115)	Hypothetical protein *Bacillus sonorensis* WP_006640209.1 (333) / Phage Integrase family protein *Bacillus subtilis* WP_004399568.1 (333) / INT_REC_C domain	84/91 (272/296) 82/91 (266/298)
29	67	+(25175-25336)	Hypothetical protein *Reinekea blandensis* WP_008045292.1 (242)	41/64 (16/25)
30	203	+(25661-26269)	Hypothetical protein *Anoxybacillus sp.* SK3-4 WP_021094316.1 (196) / Sortase *Lactobacillus gasseri* WP_003652011.1 (176)	38/56 (72/108) 32/49 (56/108)
31	54	+(26256-26471)	Hypothetical protein *Lachnobacterium bovis* WP_029067320.1 (74)	37/56 (15/23)
32	58	-(26621-26794)	Hypothetical protein Clopa_1906 *Clostridium pasteurianum* BC1 YP_007940483.1 (52) / Ribbon-helix-helix DNA binding RHH_3 domain	50/75 (26/39)
33	54	+(26795-26956)	Hypothetical protein	N/A
34	74	+(27017-27238)	Hypothetical protein *Bacillus licheniformis* WP_021837869.1 (78)	30/57 (22/42)
35	513	+(27335-28873)	Hypothetical protein *Paenibacillus elgii* WP_010498387.1 (442)	28/45 (127/203)
36	66	+(29045-29242)	Hypothetical protein *Desulfotomaculum ruminis* WP_013840758.1 (65)	39/61 (17/27)
37	65	+(29208-29402)	Ribonucleotide reductase *Clostridium straminisolvens* JCM 21531 GAE90507.1 (354)	34/55 (23/38)
38	213	+(29435-30073)	Transcriptional modulator of MazE / toxin MazF *Desulfotomaculum acetoxidans* DSM 771 YP_003193197.1 (136) / PemK superfamily	44/64 (54/80)
39	58	+(30618-30791)	Hypothetical protein *Bacillus cereus* WP_002164681.1 (56)	67/81 (36/44)
40	114	+(30799-31140)	Hypothetical protein *Bacillus cereus* WP_000073816.1 (115) / PemK-like protein *Desulfitobacterium dichloroeliminans* LMG P-21439 YP_007221476.1 (125) / PemK superfamily	60/78 (68/90) 48/64 (56/75)
41	253	+(31137-31895)	Putative phage immunity protein; phage SPbeta *Bacillus amyloliquefaciens* TA208 YP_005540881.1 (208)	36/59 (72/120)
42	530	+(31948-33537)	SPbeta phage protein *Bacillus sonorensis* WP_006640220.1 (464) / C-terminal tape_meas_TP901 domain	31/48 (159/253)
43	1914	+(33848-39589)	Lytic transglycosylase *Bacillus subtilis* WP_017696892.1 (2296) / Lytic transglycosylase and goose egg white lysozyme domains	41/59 (443/652)
44	283	+(39653-40501)	Conserved domain protein *Paenibacillus* WP_009671521.1 (284)	50/68 (139/190)
45	397	+(40513-41703)	Hypothetical protein *Paenibacillus* WP_009671508.1 (392) / Flagellin *Lactobacillus mucosae* WP_006501042.1 (680)	56/72 (221/288) 24/44 (86/160)
46	100	+(41684-41983)	Hypothetical protein *Paenibacillus* WP_009671531.1 (448)	49/68 (42/58)
47	337	+(41997-43007)	Hypothetical protein *Paenibacillus* WP_009671533.1 (335) / N-terminal GPI_anchored domain	46/70 (152/235)
48	557	+(43023-44696)	Hypothetical protein *Bacillus cereus* WP_016085028.1 (230) / Carbohydrate-binding CenC domain protein *Exiguobacterium sp.* S17 WP_016510078.1 (433)/ two CBM (carbohydrate binding module) domains	38/57 (88/133) 43/61 (66/93)

49	265	+(44683-45477)	Hypothetical protein *Brevibacillus brevis* WP_017251833.1 (204) / C-terminal DUF_4376 domain	32/50 (82/129)
50	155	+(45483-45947)	Hypothetical protein *Bacillus cereus* WP_016085021.1 (171) / N-terminal DUF_830 domain Orthopoxvirus protein of unknown function	41/62 (63/96)
51	102	+(46003-46308)	Hypothetical protein *Bacillus sp*. AP8 WP_019241528.1 (96) / Holin *Enterococcus faecium* WP_002314927.1 (80) / XhlA domain	63/81 (34/44) 38/62 (30/50)
52	82	+(46314-46559)	Holin *Paenisporosarcina sp*. TG-14 WP_017380421.1 (85) / Phage_holin superfamily domain	56/77 (40/56)
53	314	+(46578-47519)	Peptidase M15 *Bacillus megaterium* WP_016763815.1 (231)/ peptidase M15B and M15C DD-carboxypeptidase VanY/endolysin *Bacillus megaterium* WSH-002 YP_005495499.1 (231) / VanY, Peptidase_M15_4, PG_binding_1 domains	57/74 (128/166) 55/74 (122/223)
54	128	-(47783-48166)	Site-specific recombinase, DNA invertase Pin *Clostridium sp*. BNL1100 YP_005148394.1 (515) / N-terminal Zn_ribbon_recom domain	41/57 (28/35)
55	433	-(48186-49484)	Pyrimidine-nucleoside phosphorylase *Clostridium intestinale* WP_021801470.1 (432) / Glycos_transf_3, PYNP_C superfamily domains	65/80 (279/349)
56	676	-(49481-51508)	Resolvase domain-containing protein *Bacillus amyloliquefaciens* subsp. plantarum YAU B9601-Y2 YP_005420773.1 (569) / N-terminal Ser_recombinase, Zn_ribbon_recom domains	37/58 (199/309)
57	167	-(51782-52282)	Hypothetical protein *Paenibacillus polymyxa* WP_019687525.1 (173) / Putative Holliday junction resolvase *Lactococcus* phage 949 YP_004306283.1 (226) / RuvC_resolvase superfamily domain	57/73 (89/115) 40/63 (65/104)
58	175	-(52367-52891)	CtxB *Vibrio* phage CTX AF516344_3 (124) / Tyrosine kinase family protein *Microcystis aeruginosa* WP_002733350.1 (341)	19/32 (34/57) 34/57 (28/47)
59	250	-(52923-53672)	UvrD/REP helicase *Eggerthella sp.* CAG:1427 WP_021899660.1 (1084)	26/42 (38/61)
60	322	-(54011-55074)	Hypothetical protein Bacillus sp. FJAT-14578 WP_028394443.1 (361) / Contains frame shifts	N/A
61	105	-(55168-55482)	Molybdate ABC transporter, periplasmic molybdate-binding protein *Corynebacterium durum* WP_006062076.1 (222)	39/56 (20/29)
62	106	-(55516-55833)	Hypothetical protein *Bacillus cereus* WP_016085003.1 (102)	46/64 (42/59)
63	80	-(55864-56103)	Hypothetical protein *Desulfovirgula thermocuniculi* WP_027718486.1 (143)	24/53 (17/38)
64	113	-(56125-56463)	Aldehyde oxidase-like *Monodelphis domestica* XP_001379598.1 (1342)	35/56 (19/31)
65	63	-(56472-56660)	Putative uncharacterized protein *Clostridium sp.* CAG:451 WP_022469031.1 (66)	68/80 (41/48)
66	194	-(56673-57254)	Hypothetical protein *Mesorhizobium amorphae* WP_006204368.1 (187) / Ntn_hydrolase superfamily protease-like domain	47/67 (85/122)
67	203	-(57484-58092)	Hypothetical protein *Brevibacillus laterosporus* WP_018672623.1 (266)/ VirB10-like	27/50 (41/77)
68	169	-(58175-58681)	Dihydrofolate reductase *Aneurinibacillus aneurinilyticus* WP_021623827.1 (168) / DHFR superfamily domain	49/67 (81/112)
69	271	-(58682-59494)	Thymidylate synthase *Clostridium difficile* CD196 YP_003213108.1 (276) / TS_Pyrimidine_HMase domain	53/68 (145/189)
70	1366	-(59519-63622)	Putative DNA gyrase B subunit *Clostridium botulinum* D str. 16868 KEH96509.1 (1417) / contains an amber mutation	46/64 (620/875)
71	57	-(63695-63865)	Hypothetical protein CHLNCDRAFT_140300 *Chlorella variabilis* XP_005850790.1 (159)	34/55 (13/21)
72	78	-(63950-64183)	Hypothetical protein *Paenibacillus elgii* WP_010497743.1 (75)	36/60 (20/33)
73	91	-(64231-64503)	Conserved hypothetical protein *Bacillus pumilus* WP_003213202.1 (88)	48/66 (43/59)
74	96	-(64681-64968)	Hypothetical protein *Bacillus methanolicus* WP_004438575.1 (111)	61/76 (59/74)

75	55	-(65005-65169)	Hypothetical protein 0305phi8-36p083 *Bacillus* phage 0305phi8-36 YP_001429809.1 (97)	37/67 (17/31)
76	70	-(65224-65433)	Ferredoxin--NADP reductase *Acaryochloris marina* WP_012163372.1 (296)	47/62 (21/28)
77	52	-(65456-65611)	Pyrimidine-nucleoside phosphorylase *Bacillus cereus* WP_014300201.1 (78)	48/54 (16/18)
78	54	-(65656-65817)	Hypothetical protein BCQ_PT52 *Bacillus cereus* Q1 YP_002533118.1 (53) / DUF3797 domain	67/90 (34/46)
79	116	-(65851-66198)	Hypothetical protein HD73_0395 *Bacillus thuringiensis* serovar kurstaki str. HD73 YP_007419496.1 (157)	58/70 (69/84)
80	82	-(66322-66567)	Putative uncharacterized protein *Ruminococcus* sp. CAG:330 WP_022409890.1 (169)	32/51 (18/29)
81	239	-(66963-67682)	Hypothetical protein *Bacillus nealsonii* WP_016203911.1 (170) / NTP phosphohydrolase domain protein *Bacillus subtilis* subsp. subtilis str. NCIB 3610 YP_008244161.1 (174) / NTP-PPase_YP_001813558 MazG-like domain	64/77 (104/127) 61/74 (97/119)
82	84	-(67754-68005)	Thioredoxin *Bacillus sp.* BT1B_CT2 WP_009328268.1 (83) / NrdH-redoxin family domain	47/69 (37/54)
83	946	-(68046-70883)	Vitamin B12-dependent ribonucleotide reductase *Youngiibacter fragilis* WP_023388736.1 (1012) / RNR_II_dimer domain	47/63 (470/637)
84	77	-(71186-71416)	Hypothetical protein *Amycolatopsis orientalis* WP_016330646.1 (162)	33/53 (23/37)
85	135	-(71573-71977)	Hypothetical protein *Bacillus licheniformis* WP_017474291.1 (121)	34/54 (43/68)
86	76	-(72029-72256)	Hypothetical protein *Bacillus ginsengihumi* WP_025731330.1 (60)	57/71 (32/40)
87	75	-(72935-73159)	Rap GTPase activating protein domain-containing protein 1 *Strongyloides ratti* CEF70831.1 (1573)	38/56 (19/28)
88	66	-(73228-73425)	YjgP/YjgQ family permease *Cellulophaga geojensis* KL-A EWH12381.1 (468)	34/60 (20/35)
89	186	-(73443-74000)	Modification methylase CviRI *Roseburia intestinalis* WP_006855705.1 (150) / Methyltransf_26 domain	46/61 (70/94)
90	292	-(74034-74909)	Hypothetical protein *Enterococcus faecalis* WP_010785554.1 (595) / DNA-cytosine methyltransferase *Pseudoramibacter alactolyticus* WP_006598565.1 (582) / N6_N4_Mtase domain	46/64 (139/193) 44/64 (128/189)
91	113	-(75113-75451)	Hypothetical protein *Geobacillus thermoglucosidasius* WP_003253514.1 (245) / Pep_T-like domain	69/84 (59/73)
92	110	-(75461-75790)	Hypothetical protein *Caldalkalibacillus thermarum* WP_007504287.1 (100)	37/56 (27/41)
93	228	-(76126-76809)	Restriction endonuclease, partial Cannes 8 virus AGV01783.1 (516)	28/47 (21/36)
94	77	-(76806-77036)	Hypothetical protein EF87_21880 *Bacillus amyloliquefaciens* KDN88731.1 (103)	29/53 (18/34)
95	167	-(77033-77533)	SPBc2 prophage-derived protein YorR *Bacillus amyloliquefaciens* subsp. plantarum YAU B9601-Y2 YP_005421143.1 (165) / Thymidylate kinase *Methanotorris formicicus* WP_007043565.1 (185) / dNK domain	55/73 (89/119) 27/44 (53/86)
96	218	-(77759-78412)	3D domain protein *Aneurinibacillus aneurinilyticus* WP_021622015.1 (346) / 3D superfamily domain	48/65 (50/68)
97	1014	-(78425-81469)	Hypothetical protein *Bacillus cereus* WP_016094804.1 (1046) / DNA polymerase III alpha subunit *Clostridium botulinum* B str. Osaka05 BAO04764.1 (1031) / PHP_PolIIIA_DnaE3 domain	64/79 (652/808) 53/71 (538/726)
98	560	-(81495-83174)	yorK protein (Fragment) *Clostridium botulinum* B str. Osaka05 BAO04763.1 (561) / single-stranded DNA exonuclease *Bacillus vallismortis* WP_010331034.1 (576) / DHH family domain	49/67 (275/379) 46/63 (257/355)
99	349	-(83179-84225)	DNA primase *Faecalibacterium prausnitzii* SL3/3 YP_007800891.1 (340) / TOPRIM_DnaG_primases domain	40/59 (136/202)
100	170	-(84240-84749)	numod4 motif family protein *Ruminococcus* sp. CAG:9 WP_022380330.1 (241) / HNH endonuclease *Streptococcus dysgalactiae* subsp. equisimilis AC-2713 YP_006905070.1 (196) / HNH_3 domain	55/72 (88/117) 40/60 (62/93)
101	458	-(84754-86127)	Hypothetical protein *Clostridium bolteae* WP_002573415.1 (482) / Replicative DNA helicase	40/61 (190/295)

			Paenibacillus elgii WP_010497788.1 (529) / DnaB domain	41/60 (195/288)
102	126	-(86188-86565)	Hypothetical protein *Subdoligranulum* sp. 4_3_54A2FAA WP_009323218.1 (235) / phosphate uptake regulator, PhoU *Thermoplasmatales archaeon* SCGC AB-539-N05 WP_008441552.1 (232)	34/52 (40/63) 33/60 (23/42)
103	97	-(86513-86803)	Unique cartilage matrix-associated protein-like *Macaca mulatta* XP_001087282.2 (132)	28/42 (30/46)
104	384	-(86909-88060)	Hypothetical protein *Subdoligranulum* sp. 4_3_54A2FAA WP_009323219.1 (385)	36/58 (137/227)
105	335	-(88113-89117)	Hypothetical protein *Clostridium bolteae* WP_002573419.1 (355)	42/63 (144/215)
106	430	-(89499-90788)	Hypothetical protein *Bacillus cereus* WP_016094813.1 (433) / ATP-dependent DNA ligase *Paenibacillus alvei* WP_005552334.1 (430) / Adenylation_kDNA_ligase_like domain	48/68 (208/296) 48/69 (207/300)
107	74	-(90794-91015)	Hypothetical protein TCA2_4414 *Paenibacillus* sp. TCA20 GAK41922.1 (80)	73/86 (16/19)
108	276	-(91058-91885)	DNA adenine methylase family protein *Clostridium difficile* WP_021425109.1 (276) / Dam domain	61/78 (166/214)
109	85	-(91925-92179)	Aspartate carbamoyltransferase *Pandoraea* sp. B-6 WP_026131998.1 (427)	32/50 (24/37)
110	112	-(92214-92549)	Hypothetical protein *Brevibacillus brevis* WP_017248624.1 (92) / fliH domain	39/59 (30/46)
111	83	-(92772-93020)	Neurabin-1-like *Lepisosteus oculatus* XP_006636653.1 (1370)	32/55 (23/40)
112	74	-(93020-93241)	Resolvase *Salinicoccus carnicancri* WP_017549374.1 (194)	40/60 (20/30)
113	101	-(93298-93600)	Hypothetical protein BMQ_3493 *Bacillus megaterium* QM B1551 YP_003563949.1 (100)	37/48 (34/45)
114	68	-(93663-93866)	Hypothetical protein *Bacillus* phage vB_BanS-Tsamsa YP_008873365.1 (114)	42/60 (25/36)
115	151	-(93901-94353)	Hypothetical protein *Aneurinibacillus aneurinilyticus* WP_021624839.1 (146) / GIY-YIG_UvrC_Cho domain	31/55 (44/79)
116	107	-(94405-94725)	Hypothetical protein C623_0204625 *Bacillus thuringiensis* serovar aizawai str. Hu4-2 ETE99341.1 (100)	28/59 (29/61)
117	115	-(94763-95107)	Hypothetical protein *Bacillus macauensis* WP_007201879.1 (242)	33/58 (38/67)
118	84	-(95137-95388)	Hypothetical protein *Bacillus cereus* WP_001020283.1 (87)	53/74 (42/59)
119	89	-(95430-95696)	Hypothetical protein *Aneurinibacillus aneurinilyticus* WP_021621784.1 (121)	47/68 (40/58)
120	254	-(95731-96492)	Hypothetical protein BRADO3889 *Bradyrhizobium sp.* ORS 278 YP_001205874.1 (102) /	46/66 (30/43)
121	73	-(96531-96749)	PTS mannose transporter subunit IID *Clostridium novyi* B str. ATCC 27606 KEI13252.1 (139)	30/50 (24/40)
122	60	-(96773-96952)	Hypothetical protein BCP78_0087 *Bacillus* phage BCP78 YP_006907922.1 (59)	34/66 (18/35)
123	195	-(97030-97614)	Thymidine kinase *Carnobacterium* sp. AT7 WP_007720733.1 (193) / PRK04296 domain	49/66 (93/125)
124	162	-(97625-98110)	Nucleoside 2-deoxyribosyltransferase *Lysinibacillus boronitolerans* WP_016993282.1 (163) / Nuc_deoxyrib_tr domain	42/65 (68/105)
125	79	-(98132-98343)	Hypothetical protein *Bacillus* phage vB_BanS-Tsamsa YP_008873397.1 (95)	57/78 (39/54)
126	108	-(98340-98663)	Conserved protein of unknown function *Bacillus amyloliquefaciens* subsp. plantarum UCMB5033 YP_008413036.1 (124)	37/57 (38/59)
127	84	-(98680-98931)	Putative RNaseH uncultured marine crenarchaeote HF4000_ANIW97P9 ABZ07137.1 (118)	29/48 (22/37)
128	50	-(98901-99050)	Hypothetical protein	N/A
129	78	-(99116-99349)	Hypothetical protein H839_15993 *Geobacillus stearothermophilus* NUB3621 EZP75017.1 (74)	64/68 (16/17)
130	65	-(99381-99575)	Ribosome small subunit-dependent GTPase A *Acinetobacter calcoaceticus* WP_016139273.1 (353)	47/72 (17/26)

131	56	-(99608-99775)	Hypothetical protein *Bacillus flexus* WP_025909346.1 (135)	67/76 (37/42)
132	63	-(99805-99993)	XRE family transcriptional regulator Serratia sp. Ag2 KFK95694.1 (178)	34/65 (16/31)
133	82	-(100019-100264)	Putative small protein *Oscillatoriales cyanobacterium* JSC-12 WP_009556859.1 (84)	37/82 (21/30)
134	85	-(100318-100575)	Hypothetical protein V529_20360 *Bacillus amyloliquefaciens* SQR9 AHZ16062.1 (87)	31/49 (26/42)
135	155	-(100597-101061)	Ribonuclease H *Sporolactobacillus vineae* WP_010631855.1 (149) / Rnase_HI_prokaryote_like domain	67/82 (99/122)
136	124	-(101095-101466)	Hypothetical protein *Eubacterium plexicaudatum* WP_004067758.1 (107)	44/69 (27/43)
137	65	-(101478-101672)	Cobalt-precorrin-4 C(11)-methyltransferase *Thiocystis violascens* WP_014778112.1 (271)	43/71 (15/25)
138	127	-(101719-102099)	Hypothetical protein GBVE2_gp051 *Geobacillus* virus E2 YP_001285857.1 (128) / YopX family protein *Staphylococcus* phage vB_SepiS-phiIPLA5 YP_006560999.1 (129) / YopX domain	61/75 (80/99) 39/53 (52/71)
139	85	-(102130-102384)	Cytochrome P450 *Fischerella* sp. PCC 9605 WP_026734540.1 (437)	33/44 (25/33)
140	173	-(102426-102944)	AbrB family transcriptional regulator *Lentibacillus jeotgali* WP_010532485.1 (106)	66/75 (29/33)
141	71	-(102951-103163)	Hypothetical protein *Bacillus flexus* WP_025909277.1 (66) / HTH and Lar_restr_allev domains	62/70 (37/42)
142	148	-(103175-103618)	Histidine kinase *Streptomyces megaspores* WP_031506246.1 (420)	30/47 (21/33)
143	55	-(103721-103885)	Hypothetical protein	N/A
144	73	-(103872-104060)	Hypothetical protein *Brevibacillus brevis* WP_017248777.1 (65)	29/59 (18/37)
145	373	-(104319-105440)	Transposase *Desmospora* sp. 8437 WP_009710148.1 (377) / OrfB_IS605 domain	69/83 (262/314)
146	55	+(105536-105700)	Hypothetical protein *Clostridium perfringens* WP_003458470.1 (48) / RHH_3 domain	57/80 (24/34)
147	334	-(106015-107016)	Integrase *Anoxybacillus* sp. DT3-1 WP_009362130.1 (334) / INT_REC_C domain	46/68 (152/227)
148	445	-(107034-108368)	SPBc2 prophage-derived protein YopQ *Anoxybacillus* sp. DT3-1 WP_009362131.1 (445) / DndB superfamily domain	55/74 (244/329)
149	342	-(108440-109465)	Phage integrase family site specific recombinase *Staphylococcus epidermidis* RP62A YP_189166.1 (347) / INT_REC_C domain	33/54 (113/185)
150	151	-(110652-111104)	Putative uncharacterized protein *Firmicutes* bacterium CAG:449 WP_022266857.1 (141)	37/51 (47/66)
151	252	-(111149-111904)	DNA adenine methylase *Alicyclobacillus hesperidum* WP_006446198.1 (267) / Dam domain	43/62 (110/159)
152	255	-(111947-112711)	DNA-cytosine methyltransferase *Bacillus cereus* WP_000934366.1 (248) / N6_N4_Mtase domain	61/77 (154/197)
153	75	-(112743-112967)	Hypothetical protein *Bacillus subtilis* WP_019712282.1 (77)	45/68 (33/50)
154	392	-(113099-114274)	Hypothetical protein *Brevibacillus brevis* WP_017248728.1 (397) / Transposase, IS605 OrfB family *Geobacillus* sp. WCH70 YP_002951068.1 (392) / OrfB_IS605 domain	63/78 (248/309) 41/61 (157/237)
155	233	-(114576-115274)	Hypothetical protein *Bacillus azotoformans* WP_003329177.1 (241) / Nucleotidyltransferase *Bacillus* phage Grass AGY47305.1 (246)	51/68 (115/154) 31/54 (73/126)
156	486	-(115424-116881)	IS transposase *Geobacillus* sp. WCH70 YP_002948967.1 (487) / OrfB_IS605 domain	75/87 (365/428)
157	134	-(116896-117297)	Transposase *Anoxybacillus* sp. SK3-4 WP_021094952.1 (133) / Y1_Tnp domain	84/95 (112/127)
158	76	-(117404-117559)	Hypothetical protein *Bacillales* WP_015252758.1 (172)	28/60 (16/35)
159	84	-(117643-117894)	Hypothetical protein *Anoxybacillus* sp. SK3-4 WP_021095442.1 (83)	87/92 (72/77)
160	63	-(117932-118120)	Flagellar protein FliT *Virgibacillus halodenitrificans* CDQ30829.1 (117)	38/64 (20/34)

161	54	-(118193-118354)	Uncharacterized protein LOC100785018 *Glycine max* XP_006599955.1 (337)	44/66 (16/24)
162	91	-(118392-118664)	Hypothetical protein *Shigella* phage Shf125875 AIM50726.1 (120)	41/57 (22/31)
163	51	-(118797-118952)	Hypothetical protein CPR_C0019 *Clostridium* phage phiSM101 YP_699948.1 (44)	32/75 (13/31)
164	71	-(119206-119418)	Hypothetical protein *Amycolatopsis azurea* WP_005166724.1 (271) / Orthopox_35kD domain	36/58 (18/29)
165	49	-(119387-119533)	Hypothetical protein DJ51_5110 *Bacillus cereus* KFL86206.1 (39)	61/75 (22/27)
166	49	-(119538-119684)	Hypothetical protein JCM16418_5101 *Paenibacillus pini* JCM 16418 GAF10872.1 (75)	47/68 (21/31)
167	82	-(119614-119856)	Hypothetical protein *Kineococcus radiotolerans* WP_011981686.1 (85)	29/56 (16/31)
168	52	-(119890-120045)	Hypothetical protein CRE_08251 *Caenorhabditis remanei* XP_003109456.1 (241)	47/61 (16/21)
169	76	-(120076-120303)	Hypothetical protein *Bacillus thuringiensis* WP_030030167.1 (109)	49/73 (37/55)
170	61	-(120344-120526)	Hypothetical protein *Catenulispora acidiphila* WP_012787201.1 (61)	42/50 (25/30)
171	54	-(120526-120678)	CRISPR-associated protein Csm1 *Thermococcus onnurineus* WP_012571853.1 (777) / Cas10_III domain	38/55 (17/25)
172	83	-(120747-120986)	Hypothetical protein *Paenibacillus elgii* WP_010499937.1 (86)	36/65 (28/51)
173	382	-(121143-122288)	Hypothetical protein *Paenibacillus polymyxa* WP_019687721.1 (340)	25/44 (55/96)
174	320	-(122590-123549)	Hypothetical protein GWCH70_2831 *Geobacillus* sp. WCH70 YP_002950780.1 (321) / HTH_36 domain	65/78 (211/254)
175	65	-(123612-123806)	Hypothetical protein Bsph_1942 *Lysinibacillus sphaericus* C3-41 YP_001697661.1 (61)	52/80 (32/49)
176	212	-(124164-124799)	Hypothetical protein *Staphylococcus aureus* WP_016187599.1 (226) / Beta_clamp superfamily domain	34/51 (41/63)
177	94	-(124842-125123)	Hypothetical protein *Alicyclobacillus contaminans* WP_026973934.1 (159)	57/71 (20/25)
178	390	-(125137-126306)	Hypothetical protein *Paenibacillus elgii* WP_010499918.1 (376) / COG6 domain	38/63 (76/126)
179	61	-(126354-126536)	Hypothetical protein *Bacteroides* WP_004325856.1 (92)	35/58 (16/27)
180	62	-(126559-126744)	Hypothetical protein *Desulfotomaculum kuznetsovii* WP_013823457.1 (116)	39/61 (14/22)
181	55	-(126772-126936)	Hypothetical protein IscW_ISCW014631 *Ixodes scapularis* XP_002414485.1 (369) / COG3264 domain	41/56 (19/26)
182	60	-(126990-127169)	Hypothetical protein	N/A
183	88	-(127240-127461)	Hypothetical protein CANTEDRAFT_116679 *Candida tenuis* ATCC 10573 XP_006689815.1 (236)	38/57 (17/26)
184	151	-(127652-128104)	Appr-1-p processing protein *Paenibacillus lactis* WP_007128434.1 (149) / Macro_Poa1p_like domain	58/74 (87/111)
185	281	-(128253-129095)	Hypothetical protein *Clostridium botulinum* WP_003374316.1 (254) / toxin-antitoxin system, toxin component, Bro family *Clostridium hathewayi* WP_006775691.1 (279) / ORF6N domain	49/67 (59/82) 40/60 (60/90)
186	1048	-(129429-132572)	Hypothetical protein *Bacillus licheniformis* WP_017474472.1 (644)	43/65 (111/170)
187	59	-(132699-132875)	Hypothetical protein *Desulfotomaculum alcoholivorax* WP_027363725.1 (79)	30/48 (15/24)
188	64	-(132673-132864)	Hypothetical protein *Acinetobacter junii* WP_004954691.1 (420)	42/60 (16/23)
189	70	-(133553-133762)	Arginyl-tRNA synthetase *Pseudomonas syringae* KFE56075.1 (578)	38/56 (18/27)
190	61	-(133979-134161)	Uncharacterized protein LOC101122157 *Ovis aries* XP_004003698.1 (250)	30/60 (18/36)
191	98	-(134262-134555)	Transitional endoplasmic reticulum ATPase TER94-like *Apis florea* XP_003692162.1 (893)	32/47 (23/34)
192	103	-(134657-134929)	Luciferase *Mycobacterium* sp. 360MFTsu5.1 WP_029105512.1 (282)	36/55 (20/31)

193	67	-(135051-135251)	Unknown protein Candidatus *Kuenenia stuttgartiensis* CAJ74213.1 (66)	38/49 (21/27)
194	70	-(135354-135563)	Uncharacterized lipoprotein ymbA *Xenorhabdus poinarii* G6 CDG21808.1 (206)	32/50 (17/27)
195	172	-(135584-136099)	5 nucleotidase deoxy cytosolic type C Firmicutes bacterium CAG:582 WP_022178382.1 (208)	36/58 (26/42)
196	199	-(136158-136754)	Hypothetical protein BAUCODRAFT_38792 *Baudoinia compniacensis* UAMH 10762 EMC91680.1 (339)	33/49 (24/36)
197	212	-(136824-137459)	Hypothetical protein *Brevibacillus brevis* WP_017248816.1 (219)	45/63 (98/138)
198	121	-(137538-137900)	Hypothetical protein *Bacillus cereus* WP_000787323.1 (75)	46/68 (31/46)
199	68	-(137981-138184)	ATP-binding cassette sub-family A member 3-like *Trichechus manatus latirostris* XP_004373510.1 (1758)	33/58 (22/39)
200	59	-(138229-138405)	DNA topoisomerase I *Bacteroides uniformis* WP_005834124.1 (715)	45/55 (22/27)
201	372	-(139086-140201)	Hypothetical protein CTC01563 *Clostridium tetani* E88 NP_782174.1 (419)	43/56 (165/217)
202	179	-(140391-140927)	Phosphate ABC transporter permease *Thioalkalivibrio* sp. ARh3 WP_018864092.1 (555)	26/54 (22/46)

Table S2

Bacterial strain / note	CRISPR spacer sequence	CRISPR repeat unit sequence	Location on GVE3
***G. thermoglucosidasius* -11955**			
	(C)ATTCAACAACAGG[G/A]GAA[G/-]AAAAAGA[A/C]CTTCACACAA	GTTTCAATTCCTTATAGGTAAGATACAAAC	Intergenic 135200bp
	(GTGG)GATGGC[G/A]AC[C/T][T/G][A/C][GCG/TAC]CTGATGATGGCAAGCC(GA)	GTTTCAATTCCTTATAGGTAAGATACAAAC	Intergenic 138460bp
***G. thermoglucosidasius* -C56-YS93**	AACAA[G/T]CGCAAAGGTTT[A/C][A/CG][A/-]TTTTCCTTTTT[C/T]AA[G/A]CGT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF184
	ATAC[T/C][A/A][G/A]ACTTTCTTT[G/A]TATTGTGCGTATGGCTCGT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF195
	GCCAAATTTTTTATCTATCCAAGAGTAGCACCTT(TCC)	GTTTGTATCTTACCTATGAGGAATTGAAAC	Intergenic 139000bp
	TCGACATCAGGAATTTCGTCGATAAATACTTTGAA(TAT)	GTTTGTATCTTACCTATGAGGAATTGAAAC	Intergenic 134700bp
	TGAGAACATAAGCGAATTTTCCATTCGAG[A/C][A/A][T/A]ATT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Start of ORF8
	(TAA)TAA[T/C]TGTAAAATCT[AC/GT]TTTAATACTGGT(GCGCC)	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF196
	TTATATCTTC[GC/AT][T/G][A/T]TTAAAG[T/C]AGTCATGCCATCTG	GTTTCAATTCCTTATAGGTAAGATAAAACC	Inside ORF196
	(TA)TA[T/C]TTTGC[G/A]CAAATGAATACAATTGAA[T/C][C/T][A/A]ATTGG(G)	GTTTCAATTCTTCATAGGTAAGATAAAAAC	Inside ORF146
	CTTTTAG[C/T]TTCATATTGCTTGA[G/A]CCACG[A/G]AG[A/C]GAAGT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF69
	GCTCA[T/A][A/AAC][TTT/C]GCC[TT/CA]ATTTTGCGTTCTAGATG[A/G]CTTCC	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic 133400bp
	AGCGTCTG[GG/AA]AGCGTGTGAAG[T/G]TGATAGG[A/T]AAAG(GA)	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic 133420bp

	(A)TG[G/A]TGAAA[C/A]GAAA[AATA/GGAG]TTGTGAGAAGAGT(CTTGA)	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF4
	TATGATCCTCCCTTTTCTGTACAATACCTTAAACTT	GTTTTATCTGAACGTAGTGGGATATAAAG	Start of ORF32
	ATGAC[T/C]GC[TG/AA]TTGATTGGAAAGC[A/C]GG[C/T]AT[T/C]CCAAA(TG)	GTTTTATCTGAACGTAGTGGGATATAAAG	ORF42
	AAAGAAGCTTT[G/A]CAAGAATATATTGGAAAAATGGA	GTTTTATCTGAACGTAGTGGGATATAAAG	Inside ORF40
***G. thermoglucosidasius* - TNO-09.020**	GCGTGTGAAGTGG[G/A]TAGGT[G/-]AA[G/A]GA[G/T]AAAA[C/T]AAAA	GTTTCAATTCCTTATAGGTAAGATACAAAC	Intergenic 133420bp
	G[G/A]ACACCTGCAACCTAACTAAAT[A/-]AAA[C/T]GAATGGAGGAA	GTTTCAATTCCTCATAGGTAAGATACAAAC	Intergenic, beginning of ORF193
	AT[A/C]GCATCCCAACGATTATCATCACCACTATAAGT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF44
	CTACATACTTTTTGTGACTCCATGACTT[T/C][TA/GG]CGTT	GTTTCAATTCCTCATAGGTAAGATACAAAC	Inside ORF193
***G. toebii* - WCH70**	ATACTGG[CG/TA]CTCCACCGTTATCC[AT/--][A/C]TATTTTTGT	GTTTTATCTTACCTATGAGGAATTGAAAC	Inside ORF196
	CCATCATCAGGTAGCAAGTTGCCATCTTGCTACGACAAG	GTTTGTATCTTAACTATGAGGAATTGAAAC	Intergenic 138330
	TTGACAGGATATTGACCAAGCTCACCCCGTCTGCCCG	CTTTATATCCCACTACGTTCAGATAAAAC	End of ORF 143
***G. thermoglucosidasius* - Y4.1MC1**	GGATTAGTTGGC[G/A][C/TAG]---TG[T/G]TTTAG[G/T]AC[C/A]ATT	GTTTGTATCTTACCTATGAGGAATTGAAAC	Inside ORF38
	[T/C]TTTTCACTAAT[A/G]AA[A/G]AAGTCTGGATAGGATTGTTG	GTTTCAATTTTCCTTATAGGTAAGATAAAAAC	Inside ORF50
	GTCGAATGACGAACG[A-/CC]AGTGAGGAATGAGACAAGCA	GTTTCAATTTTCCTTATAGGTAAGATAAAAAC	Intergenic 138960

	(AT)GGTGGAGT$^{G}_{A}$CCAGTATTAAA$^{G}_{A}$CAGAT$^{AC}_{TT}$TACA(AT)	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF196
	TT$^{T}_{C}$CCGTGTAT$^{C}_{G}$G$^{CG}_{GA}$TTATTTACATATGCACGAAACT	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF136
	GAATTTGGAAAATCAAGAGCGCAATA$^{T}_{C}$TCAGCAGA	GTTTCAATTCCTCATAGGATACAAAC	Inside ORF95

- Subscript sequence corresponds to phage DNA sequence

Table S3

Bacterium / Phage	Areas of similarity (length in bp)		Function encoded on GVE3	Percentage identity/gaps
	GVE3	Bacterium / Phage		
***Geobacillus toebii* WCH70**	115262-115412 (150)	2854467-2854315 (152)*	Sequence associated with IS elements	96/0
		2218799-2218949 (150)		96/0
		666917-666768 (149)		96/0
		965841-965692 (149)		96/0
		1675325-1675474 (149)		96/0
		1799931-1800080 (149)*		96/0
		1997961-1998110 (149)*		96/0
		2036474-2036623 (149)		96/0
		2699836-2699687 (149)		96/0
		2930818-2930669 (149)		96/0
		1258531-1258682 (151)*		94/1
		882630-882784 (154)*		87/0
		1582295-1582146 (149)		96/0
		1541709-1541560 (149)		96/0
		1427576-1427426 (150)*		96/0
		1675325-1675474 (149)		96/0
		537200-537048 (152)*		95/0
	115261-117428 (2167)	884159-886181 (2022)	IS elements	77/0
		896311-894137 (2174)		74/0
		1988484-1989797 (1313)		77/0
		2069551-2071616 (2065)		72/2
	115262-115827 (565)	506720-506156 (564)	IS element	90/0

	122344-123507 (1163)	2863055-2864228 (1173)	HTH containing ORF	72/5
	48714-49484 (770)	2306426-237196 (770)	N-terminal of Pyrimidine nucleoside phosphorylase	70/0
	5914-6093 (179)	2219251-2219072 (179)	IHF-like protein	81/0
	5499-5698 (199)	2869300-2869094 (206)	C-terminal of cons. hypo. protein	78/3
	104247-104483 (236)	2150056-2150293 (237)	C-terminal of ORF100 (IS605) and sequence downstream of it	73/0
	138448-138486 (38)	360836-360874 (38)	Upstream of four conserved hypothetical proteins	100/0
	104298-104334 (36)	2061807-2061771 (36)	Sequence downstream of ORF100	100/0
***Clostridium botulinum* B str. Osaka05 DNA, contig: Osaka05p1_contig002, extrachromosome 1**	78895-82892 (3997)	19479-23480 (4001)	DNA pol III alpha subunit and Single stranded exonuclease	65/4
***Bacillus* phage vB_BanS-Tsamsa**	79026-81468 (2442)	97875-100338 (2463)	DNA pol III alpha subunit	69/2
	59838-62400 (2562)	61548-64155 (2607)	DNA gyrase A and B	67/4
	33358-33413 (55)	127875-127930 (55)	Tape measure protein	80/0
***Bacillus* phage SPBc2**	24098-25107 (1009)	40307-41318 (1011)	Integrase-like	75/0
	36326-36985 (659)	32687-33350 (663)	Tape measure protein	71/4
	5927-6090 (163)	60843-61007 (164)	IHF-like protein	77/2
	82724-83165 (441)	105199-105640 (441)	N-terminal of Single stranded exonuclease	64/4
***Clostridium* phage c-st**	115428-116985 (1557)	4081-5668 (1587)	IS element	70/3
		16518-18105 (1587)		70/3
		158521-160108 (1587)		70/3
	115615-116519 (904)	168684-169567 (883)	IS element	69/6
	79711-81489 (1778)	51559-53317 (1758)	DNA pol III alpha subunit	66/3
	78922-79273 (351)	53761-54112 (351)	DNA pol III alpha subunit	68/2

	60218-60300 (82)	90017-89935 (82)	DNA gyrase	76/0

Table S4. PHAST annotation of three regions of phage GVE3

Region 1, total : 35 CDS.

#	CDS_POSITION	BLAST_HIT	E-VALUE
1	3313..3495	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p064; PP_00004; phage(gi9630189)	3e-06
2	3495..3755	PHAGE_Staphy_Twort_NC_007021: ORF137; PP_00005; phage(gi66391382)	4e-08
3	3947..5647	hypothetical protein SERP1620 [Staphylococcus epidermidis RP62A] gi|57867549|ref|YP_189185.1|; PP_00006	4e-56
4	5354..5365	attL ATTCGGGATATG	0.0
5	5821..6093	PHAGE_Bacill_SPBc2_NC_001884: histone-like prokaryotic DNA-binding protein family; PP_00007; phage(gi9630187)	7e-29
6	6164..7357	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00008; phage(gi564292570)	1e-72
7	7612..7803	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p061; PP_00009; phage(gi9630186)	7e-14
8	7816..9036	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p060; PP_00010; phage(gi9630185)	2e-68
9	9110..9406	hypothetical; PP_00011	0.0
10	9589..9780	phage protein [Bacillus licheniformis DSM 13 = ATCC 14580] gi|404488823|ref|YP_006712929.1|; PP_00012	3e-05
11	9830..11401	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p027; PP_00013; phage(gi9630152)	2e-27
12	11444..12346	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p057; PP_00014; phage(gi9630182)	2e-13
13	12339..14102	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p056; PP_00015; phage(gi9630181)	6e-58

14	14133..15638	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p055; PP_00016; phage(gi9630180)	2e-34
15	15649..15771	hypothetical; PP_00017	0.0
16	15803..17227	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p054; PP_00018; phage(gi9630179)	2e-23
17	17280..17714	hypothetical protein CEA_G1142 [Clostridium acetobutylicum EA 2018] gi|384457855|ref|YP_005670275.1|; PP_00019	3e-13
18	17751..18731	PHAGE_Bacill_Slash_NC_022774: hypothetical protein; PP_00020; phage(gi609217106)	6e-10
19	18801..19229	hypothetical protein CEA_G1140 [Clostridium acetobutylicum EA 2018] gi|384457853|ref|YP_005670273.1|; PP_00021	1e-05
20	19244..19657	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00022; phage(gi564292646)	3e-15
21	19667..20326	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p048; PP_00023; phage(gi9630173)	3e-06
22	20341..20652	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00024; phage(gi564292696)	7e-09
23	20649..21122	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p047; PP_00025; phage(gi9630172)	3e-06
24	21122..21838	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p046; PP_00026; phage(gi9630171)	5e-05
25	21857..22606	PHAGE_Lactoc_949_NC_015263: putative phage structural protein; PP_00027; phage(gi327197979)	2e-35
26	22670..23083	hypothetical; PP_00028	0.0
27	23143..23667	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p038; PP_00029; phage(gi9630163)	1e-06
28	23668..24093	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p037; PP_00030; phage(gi9630162)	1e-21
29	23960..23971	attR ATTCGGGATATG	0.0
30	24114..25115	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p036; PP_00031; phage(gi9630161)	2e-164
31	25136..25336	hypothetical; PP_00032	0.0

32	25517..25639	hypothetical; PP_00033	0.0
33	25661..26269	hypothetical protein Tresu_1654 [Treponema succinifaciens DSM 2489] gi|328948512|ref|YP_004365849.1|; PP_00034	2e-16
34	26256..26471	hypothetical; PP_00035	0.0
35	complement(26621..26794)	hypothetical protein Clopa_1906 [Clostridium pasteurianum BC1] gi|488770689|ref|YP_007940483.1|; PP_00036	2e-06
36	27017..27238	hypothetical; PP_00037	0.0
37	27350..28873	PHAGE_Staphy_CNPH82_NC_008722: conserved phage protein; PP_00038; phage(gi119953709)	5e-07

Region 2, total : 83 CDS.

#	CDS_POSITION	BLAST_HIT	E-VALUE
1	complement(56673..57254)	PHAGE_Vibrio_12B8_NC_021073: hypothetical protein; PP_00075; phage(gi481019685)	2e-21
2	complement(57303..57482)	hypothetical; PP_00076	0.0
3	complement(57484..58092)	hypothetical protein ERIC2_c26990 [Paenibacillus larvae subsp. larvae DSM 25430] gi|568264958|ref|YP_008968433.1|; PP_00077	1e-11
4	complement(58175..58681)	PHAGE_Bacill_phiAGATE_NC_020081: putative dihydrofolate reductase; PP_00078; phage(gi448260875)	1e-35
5	complement(58682..59494)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00079; phage(gi564292594)	2e-45
6	complement(59519..61651)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00080; phage(gi564292556)	0.0
7	complement(61664..63622)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00081; phage(gi564292558)	4e-160
8	complement(63695..63865)	hypothetical; PP_00082	0.0
9	complement(63950..64183)	hypothetical; PP_00083	0.0

10	complement(64231..64503)	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p155; PP_00084; phage(gi9630280)	8e-15
11	complement(64524..64679)	hypothetical; PP_00085	0.0
12	complement(64681..64968)	PHAGE_Bacill_SP10_NC_019487: hypothetical protein; PP_00086; phage(gi418489661)	4e-15
13	complement(65005..65160)	hypothetical; PP_00087	0.0
14	complement(65224..65388)	hypothetical; PP_00088	0.0
15	complement(65456..65593)	hypothetical; PP_00089	0.0
16	complement(65656..65817)	PHAGE_Clostr_CDMH1_NC_024144: conserved hypothetical protein; PP_00090; phage(gi640884924)	2e-07
17	complement(65851..66198)	PHAGE_Strept_phiBHN167_NC_022791: phage protein; PP_00091; phage(gi557745672)	1e-15
18	complement(66322..66567)	hypothetical; PP_00092	0.0
19	complement(66602..66856)	hypothetical; PP_00093	0.0
20	complement(66963..67682)	PHAGE_Cyanop_NATL1A_7_NC_016658: gp32; PP_00094; phage(gi372217788)	6e-08
21	complement(67754..68005)	PHAGE_Bacill_SPBc2_NC_001884: thioredoxin; PP_00095; phage(gi9630289)	2e-14
22	complement(68046..70883)	PHAGE_Halovi_HVTV_1_NC_020158: ribonucleotide reductase alpha subunit; PP_00096; phage(gi443404588)	2e-38
23	complement(70913..71161)	hypothetical; PP_00097	0.0
24	complement(71186..71416)	hypothetical; PP_00098	0.0
25	complement(71573..71977)	PHAGE_Bacill_BCP78_NC_018860: hypothetical protein; PP_00099; phage(gi410492830)	4e-13
26	complement(72029..72214)	hypothetical; PP_00100	0.0
27	complement(72405..72572)	hypothetical; PP_00101	0.0

28	complement(72606..72857)	hypothetical; PP_00102	0.0
29	complement(72935..73159)	hypothetical; PP_00103	0.0
30	complement(73228..73416)	hypothetical; PP_00104	0.0
31	complement(73443..74000)	PHAGE_Cronob_CR9_NC_023717: putative DNA methyltransferase; PP_00105; phage(gi593777337)	1e-18
32	complement(74034..74909)	PHAGE_Cellul_phi12:1_NC_021791: DNA methylase; PP_00106; phage(gi526177136)	2e-55
33	complement(74906..75094)	hypothetical; PP_00107	0.0
34	complement(75113..75451)	hypothetical; PP_00108	0.0
35	complement(75461..75790)	hypothetical; PP_00109	0.0
36	complement(75910..76077)	PHAGE_Halovi_HCTV_1_NC_021330: hypothetical protein; PP_00110; phage(gi509140762)	1e-06
37	complement(76126..76809)	hypothetical; PP_00111	0.0
38	complement(76806..77036)	hypothetical; PP_00112	0.0
39	complement(77033..77533)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00113; phage(gi564292628)	1e-24
40	complement(77759..78412)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00114; phage(gi564292603)	9e-08
41	complement(78425..81469)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00115; phage(gi564292551)	0.0
42	complement(81495..83174)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00116; phage(gi564292561)	2e-52
43	complement(83179..84225)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00117; phage(gi564292575)	4e-39
44	complement(84240..84749)	PHAGE_Xantho_Xp10_NC_004902: endonuclease of the HNH family with predicted DNA-binding module at C-terminus; PP_00118; phage(gi32128470)	3e-24

45	complement(84754..86175)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00119; phage(gi564292562)	2e-12
46	complement(86188..86562)	hypothetical; PP_00120	0.0
47	complement(86513..86848)	hypothetical; PP_00121	0.0
48	complement(86909..88060)	PHAGE_Clostr_c_st_NC_007581: hypothetical protein CST056; PP_00122; phage(gi80159742)	1e-09
49	complement(88113..89117)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00123; phage(gi564292571)	5e-18
50	complement(89499..90782)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: DNA ligase, ATP-dependent; PP_00124; phage(gi564292567)	1e-74
51	complement(90794..91015)	hypothetical; PP_00125	0.0
52	complement(91058..91885)	PHAGE_Parame_bursaria_Chlorella_virus_NY2A_NC_009898: hypothetical protein NY2A_B774R; PP_00126; phage(gi157953078)	3e-29
53	complement(91925..92164)	hypothetical; PP_00127	0.0
54	complement(92214..92549)	hypothetical; PP_00128	0.0
55	complement(92578..92745)	hypothetical; PP_00129	0.0
56	complement(92772..93020)	hypothetical; PP_00130	0.0
57	complement(93020..93241)	hypothetical; PP_00131	0.0
58	complement(93298..93600)	hypothetical protein BMD_3488 [Bacillus megaterium DSM 319] gi|295705603|ref|YP_003598678.1|; PP_00132	2e-07
59	complement(93663..93860)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00133; phage(gi564292684)	2e-06
60	complement(93901..94323)	hypothetical; PP_00134	0.0
61	complement(94405..94725)	PHAGE_Bacill_Spock_NC_022763: hypothetical protein; PP_00135; phage(gi568190861)	5e-05
62	complement(94763..95107)	hypothetical; PP_00136	0.0

63	complement(95137..95388)	hypothetical; PP_00137	0.0
64	complement(95430..95696)	hypothetical; PP_00138	0.0
65	complement(95731..96492)	PHAGE_Acanth_mimivirus_NC_014649: hypothetical protein; PP_00139; phage(gi311978204)	4e-08
66	complement(96531..96749)	hypothetical; PP_00140	0.0
67	complement(96773..96919)	hypothetical; PP_00141	0.0
68	complement(97030..97614)	PHAGE_Entero_vB_EcoM_VR7_NC_014792: Tk thymidine kinase; PP_00142; phage(gi314121676)	9e-30
69	complement(97625..98110)	PHAGE_Lactoc_949_NC_015263: putative nucleoside-2-deoxyribosyltransferase; PP_00143; phage(gi327197942)	1e-23
70	complement(98107..98343)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00144; phage(gi564292710)	1e-18
71	complement(98340..98663)	PHAGE_Bacill_SPBc2_NC_001884: hypothetical protein SPBc2p126; PP_00145; phage(gi9630251)	2e-10
72	complement(98680..98931)	hypothetical; PP_00146	0.0
73	complement(98901..99050)	hypothetical; PP_00147	0.0
74	complement(99116..99349)	hypothetical; PP_00148	0.0
75	complement(99381..99575)	hypothetical; PP_00149	0.0
76	complement(99608..99775)	hypothetical; PP_00150	0.0
77	complement(99805..99993)	hypothetical; PP_00151	0.0
78	complement(100019..100264)	hypothetical; PP_00152	0.0
79	complement(100318..100575)	hypothetical; PP_00153	0.0
80	complement(100597..101061)	PHAGE_Ostreo_OlV1_NC_014766: hypothetical protein; PP_00154; phage(gi313844138)	2e-26

81	complement(101095..101466)	hypothetical; PP_00155	0.0
82	complement(101478..101672)	hypothetical; PP_00156	0.0
83	complement(101719..102099)	PHAGE_Geobac_virus_E2_NC_009552: hypothetical protein GBVE2_gp051; PP_00157; phage(gi148747778)	2e-42

Region 3, total : 22 CDS.

#	CDS_POSITION	BLAST_HIT	E-VALUE
1	89396..89412	attL TTTTATATTTTATTTAA	0.0
2	complement(104319..105440)	PHAGE_Clostr_c_st_NC_007581: putative IS transposase (OrfB); PP_00163; phage(gi80159731)	4e-99
3	105536..105700	hypothetical protein [Acetohalobium arabaticum DSM 5501] gi|302391636|ref|YP_003827456.1|; PP_00164	8e-05
4	complement(106015..107016)	PHAGE_Clostr_c_st_NC_007581: conserved hypothetical phage-related protein; PP_00165; phage(gi80159716)	2e-21
5	complement(107034..108377)	PHAGE_Clostr_c_st_NC_007581: conserved hypothetical phage-related protein; PP_00166; phage(gi80159715)	3e-15
6	complement(108440..109465)	PHAGE_Lactoc_phiL47_NC_023574: putative integrase-recombinase; PP_00167; phage(gi589890760)	2e-31
7	complement(109487..109648)	hypothetical; PP_00168	0.0
8	complement(109725..109907)	hypothetical; PP_00169	0.0
9	complement(110501..110674)	hypothetical; PP_00170	0.0
10	complement(110652..111104)	PHAGE_Strept_K13_NC_024357: phage protein; PP_00171; phage(gi658307253)	2e-05
11	complement(111149..111904)	PHAGE_Clostr_phiSM101_NC_008265: putative modification methylase dpniia; PP_00172; phage(gi110804053)	7e-52
12	complement(111947..112711)	PHAGE_Geobac_GBK2_NC_023612: DNA methylase; PP_00173; phage(gi589893811)	5e-82

13	complement(112743..112967)	hypothetical; PP_00174	0.0
14	complement(113099..114274)	PHAGE_Staphy_vB_SauM_Remus_NC_022090: transposase; PP_00175; phage(gi530787614)	1e-11
15	complement(114352..114498)	hypothetical; PP_00176	0.0
16	complement(114576..115274)	PHAGE_Bacill_vB_BanS_Tsamsa_NC_023007: hypothetical protein; PP_00177; phage(gi564292596)	8e-29
17	complement(115424..116881)	PHAGE_Clostr_c_st_NC_007581: putative IS transposase (OrfB); PP_00178; phage(gi80159857)	6e-175
18	complement(116896..117297)	PHAGE_Clostr_c_st_NC_007581: putative IS transposase (OrfA); PP_00179; phage(gi80159868)	2e-26
19	complement(117386..117598)	hypothetical; PP_00180	0.0
20	complement(117643..117894)	hypothetical; PP_00181	0.0
21	complement(117932..118120)	hypothetical; PP_00182	0.0
22	complement(118193..118354)	hypothetical; PP_00183	0.0
23	complement(118392..118664)	PHAGE_Staphy_GH15_NC_019448: hypothetical protein; PP_00184; phage(gi418488124)	2e-05
24	127541..127557	attR TTTTATATTTTATTTAA	0.0

Table S5

Enzyme- recognition sequence (digestion detected Y/N)	Number of sites	Fragment sizes expected
*Nde*I – CATATG (N)	85	Too numerous, evenly spread over genome
*Sph*I – GCATGC (N)	6	4, 2535 (if linear), 3754, 11669, 12321, 40638, 70377 (if linear)
*Bst*EII – GGTNACC (N)	4	3278, 13846, 33925 (if linear), 35927, 54322 (if linear)
*Bgl*II – AGATCT (Y)	8	1125, 1836, 2543 (linear), 3260, 8546, 15880, 18533 (if linear), 28753, 60822
*Dra*III – GACNNNGTG (N)	6	600 (if linear), 7300, 13782, 23155 (if linear), 27163, 34442, 34856
*Sma*I – CCCGGG (N)	1	55122 and 86176 (if linear)
*Eco*RI – GAATTC (Y)	27	Too numerous, evenly spread over the genome
*Eco*RV – GATATC (N)	12	498, 1099, 1573, 2887, 4125, 5147, 10180 (if linear), 12539, 14904, 16207, 20493, 21079, 30747 (if linear)
*Pvu*II – CAGCTG (N)	2	4782bp, 136516bp (circular) or 34452 (if linear,) 102064 (if linear)
*Hind*III – AAGCTT (N)	23	Too numerous, evenly spread over genome
*Rsa*I – GTAC (N)	228	Too numerous, evenly spread over the genome
*Alu*I – AGCT (Y)	345	Too numerous, evenly spread over the genome
*Hae*III – GGCC (N)	11	Eight small fragments < 200bp, 36918 (if linear), 104380 (if linear)