Viruses of the Thermotogota; the vehicles on the highways of gene sharing.

Running title: Proviruses from Thermotogota.

Thomas Haverkamp[a#], Julien Lossouarn [b], Olga Zhaxybayeva[c], Jie Lyu[d], Nadège Bienvenu[d], Claire Geslin[d], Camilla L. Nesbø[e,f]

[a] Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, Norway.

[b] Université Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350, Jouy-en-Josas, France

[c] Department of Biological Sciences, Dartmouth College, Hanover, NH, USA.

[d] Univ Brest, CNRS, IFREMER, Laboratoire de Microbiologie des Environnements Extrêmes, F-29280 Plouzané, France

[e]Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

[f]Biozone, Department of Chemical Engineering and Applied Chemistry and BioZone, University of Toronto, 200 College Street, Toronto, Ontario, Canada, M5S 3E5

[#]present address: Department of Epidemiology, Norwegian Veterinary Institute, Oslo, Norway.

Abstract

Viruses are drivers of microbial ecology and evolution controlling populations and disseminating DNA laterally among the cells they infect. Phylogenetic and comparative genomic analyses of Thermotogota bacteria have shown high levels of lateral gene transfer with distantly related organisms, particularly with Firmicutes. One likely source of such DNA transfers is viruses, however, to date only three temperate viruses infecting *Marinitoga* bacteria have been characterized in this phylum. Here we use a bioinformatic approach to identify an additional 17 proviruses integrated into genomes of eight Thermotogota genera including *Marinitoga*. We also induce viral particle production from one of the newly identified proviruses. The proviruses fall into two groups based on sequence similarities, gene synteny and virus classification tools. Group-1 shows similar genome structure to the three previously identified *Marinitoga* viruses while Group-2 are distantly related to these viruses and have different genome organization. Both groups show close connections to Firmicutes in genomic- and phylogenetic analyses, and the Group-2 viruses show evidence of very recent transfer between these lineages and are likely capable of infecting cells from both phyla. We suggest that viruses are responsible for a large portion of the laterally transferred DNA between these distantly related lineages.

Introduction

The phylum Thermotogota comprises anaerobic fermentative bacteria, most of which are thermophiles [1]. They are found as a deep branching lineage in some universal phylogenetic trees, for example those made from 16S rRNA genes [2], and are common

in subsurface environments such as marine vents, terrestrial hot springs and deep subsurface oil reservoirs [3][4–6].

Lateral gene transfer (LGT) has been an important force shaping the genomes of Thermotogota, with Firmicutes and Archaea being their most notable gene transfer partners [1, 7, 8]. In fact Firmicutes and Thermotogota have been suggested to be linked by ‘‘highways of gene sharing’’ [8]. How these gene-sharing connections between phyla occur, is however, still unclear. Homologous recombination among closely related genomes is another important evolutionary force acting on Thermotogota genomes [4, 9]. For both the *Thermotoga* and *Mesotoga* genera, comparative genomic analyses and analyses of homologous recombination have shown that long fragments of DNA (> 10,000 nucleotides) are transferred between closely related strains [6]. Both types of gene-sharing processes might be driven by viruses.

The subsurface constitutes the largest biosphere on earth and is estimated to contain ~70% of all cells [10]. Viruses are likely to be particularly important in subsurface environments with 97% of all viruses on earth being found in soil and sediments [11, 12]. Moreover, although both prokaryotic cell and virus numbers decreases with depth, the virus to cell ratio increases with depth [11, 13, 14]. Phylogeographic studies of hyperthermophilic *Thermotoga* and mesophilic *Mesotoga* bacteria have revealed genetic interaction between geographically distant populations, particularly among the hyperthermophilic Thermotoga [4, 6]. Viruses are one potential source of such long-distance dispersal of genetic material, particularly for anaerobic organisms where surface dispersal is problematic.

However, although viruses are likely candidates for transferring DNA, evidence of viral activity is so far limited to one Thermotogota genus. Three temperate *Siphoviridae* viruses have been described that infect *Marinitoga* species [15, 16]. MPV1 infects the deep-sea marine vent bacterium *Marinitoga piezophila* and displays an intricate relationship with a plasmid co-occurring in the same host, illustrating the importance of mobile elements in these ecosystems, while MCV1 and MCV2 infect *Marinitoga camini*. The three viruses show similar genomic organization and virion morphology. Phylogenetic analyses of the virus ORFs revealed close connections to Firmicutes viruses, which prompted the conclusion that viruses belonging to these phyla sample genes from a common pool.

Here we identify 17 provirus sequences (including 13 likely complete provirus elements) in Thermotogota genomes using a bioinformatic approach that included several phage-detection tools as well as manual inspection of the results. We successfully induced one of these novel proviruses. These new proviral sequences are found in genomes from eight Thermotogota genera, including *Marinitoga*. We report that the Thermotogota proviruses fall into two distinct groups, and that the proviruses from one of these groups are very similar to Firmicutes viruses. We also hypothesize that membrane transport proteins (e.g. ABC-transporters) may serve as receptors for Thermotogota viruses. We propose a ratchet-like process, where moron genes encoding transmembrane proteins may make the host vulnerable to new viruses, which could account for the highways of gene-sharing observed between Thermotogota and Firmicutes.

Material and Methods

Data collection and genome analyses of viruses.

111 Thermotogota genomes available in Genbank or IMG [17] were accessed until June 2018. For draft genomes we combined the contigs using the "union" command from the EMBOSS package (version 6.6.0) [18]. This created 'artificially' closed-genomes needed for Prohinder (http://aclame.ulb.ac.be/Tools/Prophinder/) [19]. For consistency, we also used these 'artificially' closed genomes with the other provirus detection tools. We screened each genome for the presence of prophages using three tools. The genomes were uploaded to webservers for Prophinder and PHAST (http://phast.wishartlab.com/) [20] and processed using default settings at both websites. In addition, we screened all the genomes using PhiSpy (version 2.3) [21]. From the output we recorded: start and stop position of the provirus sequence, the size of the provirus, the number of open reading frames (ORFs), the presence of tRNA sequences. The annotation of ORFs in the predicted regions were manually screened for the presence of prophage /mobile element related genes by searching for the terms: "Viral", "Phage". Putative provirus regions with few provirus-like annotations, were discarded. This might lead to an underestimate of proviruses. In the case of the genomes composed of concatenated contigs, we discarded provirus identifications that crossed contig borders. The provirus calls for each of the three tools were compared and when proviruses were predicted by multiple tools in the same region of the chromosomes, the annotations were inspected to identify the most likely regions. When only one tool identified a putative virus sequence, we manually checked if the spanned region had to be corrected. This included the following steps: 1) PHAST can predict attR and attL integration sequences for the identified provirus, and when predicted, the region between the sites was manually inspected to make sure the

region contained virus genes. 2) We screened for the presence of a tRNA gene close to the attR/L site that could function as an integration site. 3) We determined if genes from the host genome were incorrectly included in the predicted prophage region by comparisons to closely related genomes (from the same genus) that did not have a provirus. The predicted provirus sequences were compared to publicly available genomes by searching the NCBI non-redundant (nr) database [22] with provirus ORFs as queries using BlastX ver.2.2.26 [23]. When homologs for multiple ORFs from the predicted provirus were found in the same distantly related reference genome (usually a Firmicutes genome), the identified reference genome was downloaded for a more detailed comparison. The reference genome was aligned to the host genome of the putative provirus using Progressive Mauve [24]. In most cases this visualized only the aligned proviruses in two otherwise distantly related microbial genomes and this information was included when determining the genomic coordinates of the provirus.

Proviruses were considered complete if they contain modules for lysogeny, replication, packaging, head/tail morphogenesis and lysis. If one or more of these modules were missing, the provirus was scored as partial. It should be noted that for two of the proviruses considered complete (M1135V1, M1138V1) the lysogeny module was not clearly identified and for two others (MLV1, T1074V1) the lysis module was not easily detected.

The predicted provirus sequences were compared to each other using blastn and tblastx [23] and visualized in GenoplotR [25] and Circos [26]. Provirus-ORF annotations were obtained from their respective Genbank entries, from blastp searches [23] of the NCBI non redundant (*nr*) database with an expected-value cutoff of 10^{-1}, and to the PDB

database using HHpred [27] with a probability cutoff of 99%. In addition, recombinase and terminase ORFs were annotated using InterProScan as implemented in Geneious v.10.

The provirus genomes were compared to NCBI's viral ref seq v. 94 database using VContact2 [28][29], and their morphological and taxonomic classification was assessed using Virfam [30], VIRIDIC [31] and VIPTree [32].

Possible receptor proteins for the Group 2 viruses were identified in IMG using a similarity cut-off at 50%. Homologs in *G. ferrireducens*, not available in IMG, were identified using blastp.

Phylogenetic analyses of virus genes. Each translated virus gene from the gene-family analyzed were queried using blastp against *nr* (accesses December 2019 – March 2020), as well as a local database of *Thermotogota* virus proteins, and the 20 top scoring matches were retrieved and aligned using MAFFT (expected value cut-off at 10^{-5}). Identical sequences were removed, as well as highly similar sequences from the same genus, and alignment positions with > 50% gaps were trimmed. Phylogenetic trees were constructed using RAxML [33] as implemented in Geneious v.10 with GAMMA+WAG model of substitution (with 4 rate categories) and 100 bootstrap replicates.

Virus induction and electron microscopy.

T. africanus H17ap6033 and *Petrotoga* sp. 8T1HF07.NaAc.6.1 were cultivated in a modified Ravot medium as previously described in reference [15] at 65°C or 55°C, respectively. Attempts were made to increase the viral production of the strains by using

mitomycin C, as reported previously [15, 16]. A final concentration of 5 µg/mL of mitomycin C was added to 300 mL bacterial culture at early to mid-log growth phase. After 3 hours of incubation with mitomycin C, cultures were centrifuged at 7500 rpm and 4°C for 15 min, and supernatants were ultracentrifuged at 37 000 rpm (~100 000 g) and 10°C for 1h (Beckman Optima LE-80 K; rotor 70.1.Ti). Pellets were resuspended in 100µL of buffer (10 mM Tris-HCL, 100 mM NaCl, 5 mM $CaCl_2$, 20 mM $MgCl_2$) and suspensions were prepared for negative staining electron microscopy as previously described [34]. Briefly, 5 µL of the suspensions were directly spotted onto a Formwar carbon coated copper grid. Putative virus-like particles were allowed to adsorb to the carbon layer for 2 min and excess of liquid was removed. 5 µL of a staining uranyl acetate solution (2%) was then spotted to the grid for 45 s and excess of liquid was removed again. The grid was imaged at 120 kV in a JEOL JEM 100 CXIIVR transmission electron microscope.

Results and Discussion

Genomic characterization of Thermotogota proviruses identify two distinct virus lineages.

Analysis of 111 Thermotogota genomes identified 20 proviruses, including the three already characterized viruses from *Marinitoga* and four partial proviruses (Supplementary Table S1). All available *Thermosipho melanesiensis* genomes have identical provirus sequences [35] and only one representative of these genomes were

included in our analyses below. Annotation of coding sequences (CDS) in the proviruses are reported in Supplementary Table S2.

Comparisons of the predicted provirus genomes using several different tools outlined below revealed two distinct groups of Thermotogota proviruses, here denoted as Group-1 and Group-2 (Fig. 1 - Fig. 3). For these comparative genomic analyses, we also included a provirus reported in *Marinitoga lauensis* [27] after our bioinformatic screening was performed. The genome organization differs between the two groups and the majority of top scoring blast-matches when comparing the Thermotogota proviruses occur within the two groups (Fig. 1, Supplementary Table S2). The distinction of these groups was supported by network analysis comparing the predicted provirus proteins to the NCBI virus refseq database (Supplementary Fig. S1), where the Thermotogota proviruses form two distinct clusters connected to a larger cluster mainly composed of Firmicutes viruses. This clustering was also observed by the VIPTree tool (Supplementary Fig. S2). Group-1 (N=15) is most widespread, found in the genera *Marinitoga*, *Thermosipho*, *Kosmotoga*, *Mesotoga*, *Geotoga* and *Mesoaciditoga*, while Group-2 (N=5) is limited to the genera *Petrotoga* and *Defluviitoga* (Supplementary Fig. S1). Some of the proviruses appear to be incomplete, (M1137V1, MTOLDCV1, PMV1, P9T1HF07V1), and may be in the process of being lost from the genomes or domestication.

Genomic characterization of Group-1 proviruses: None of the 15 Group-1 proviruses identified here (including the three characterized temperate viruses and the incomplete proviruses), exhibits significant nucleotide similarity with viral genomes in the NCBI virus database. Following taxonomic criteria where viruses belonging to the

same genus display > 50-70% nucleotide identity over the full genome while species display > 95% nucleotide identity [36][37], the VIRIDIC tool suggested that the complete Group-1 proviruses identified here, correspond to 13 new species and at least 11 new genera (Supplementary Table S2). Two provirus pairs, MCV1- MCV2 and M1135V1 - M1138V1, may be assigned to the same genus having intergenomic similarity scores of 56% and 60% respectively (Supplementary Table S2).

The newly identified Group-1 proviruses show the same modular structure as MPV1, MCV1 and MCV2 [15, 16] (Fig. 2). The 5' module in the provirus contains genes involved in lysogeny and is encoded on the opposite strand compared to the rest of the virus genes. This module is followed by modules for replication, packaging, morphogenesis and host lysis. As noted for the described *Marintoga* viruses [15, 16], the most variable module is the lysogeny module, where only the recombinase is conserved between different proviruses (Fig. 2). Higher level of conservation is observed for the other modules, and some of the conserved genes are discussed in more detail below. The novel proviruses identified here likely correspond to siphovirus based on their head-neck-tail gene sequences [30], which is consistent with the morphology observed for the characterized viruses in this group (MPV1, MCV1, MCV2), and TAV1 induced in the current study (see below).

All the complete Group-1 proviruses are inserted next to a tRNA gene. Eight of them (the *Marinitoga* proviruses, MPV1, MCV1, MCV2, MHV1 and M1137V2, KPV1 (*Kosmotoga pacifica*), GPV1 (*Geotoga petrae*), MLV1 (*Mesoaciditoga lauensis*)), are inserted next to tRNA-Glu, and share a very similar site-specific DNA serine recombinase (KEGG Orthology; K06400, homologs of Marpi_0291 in MPV1 from *M.*

piezophila) (Fig. 4). The *Thermosipho* proviruses also carry homologs of this recombinase but are inserted next to different tRNA genes. TMV1 and T1074V1, which have recombinases that cluster together in the tree in Fig. 4, are inserted next to tRNA-Phe. The most divergent homolog of this recombinase is found in TAV1 which is inserted next to tRNA-Pro. We suggest that these recombinases are integrases that specifically recognizes the tRNA genes the provirus is inserted next to (i.e. tRNA-Glu, tRNA-Phe or tRNA-Pro). M1135V1 and M1138V1 are inserted next to tRNA-Cys and have two unrelated ORFs with no suggested function in this position.

Terminase proteins are biological machines involved in the packaging of virus DNA into the virus particle [38]. The Group-1 virus genomes have three families of terminases (BLASTP cut-off exp. > e^{-1}, ident. > 25%, Supplementary Fig. S3). The first family, exemplified by Marpi_0320 in MPV1, has a PBSX family domain and has homologs in MCV1 and MCV2 and the proviruses in *Marinitoga* sp. 1135, *Marinitoga* sp. 1138, *Thermosipho* sp. 1074 and *M. lauensis*. The second type, exemplified by the terminase protein in TAV1 (H17ap60334_04902) has a 'Terminase_lsu_T4-like' domain and close homologs in the proviruses from *K. pacifica*, *T. melanesiensis*, *Mesotoga* sp. TolDC and *G. petrae*. Finally, the proviruses in *M. hydrogenitolerans* and *Marinitoga* sp. 1137 have a third group of similar terminases (BUA62_RS02495, LN42_01905 and LN42_00550, a pseudogene) in the same position, also with a 'Terminase_lsu_T4-like' domain, thus these are distant homologs of the second terminase family (Supplementary Fig. S3).

In addition to the recombinase and terminase, other typical virus proteins such as capsid and portal proteins were also identified (Fig. 2). Moreover, two transcription

regulators (Marpi_0297 and Marpi_0298 in MPV1), a DNA repair exonuclease (Marpi_0340 in MPV1), a single stranded DNA-binding (Marpi_0306 in MPV1) and the tail tape measure protein (Marpi_0334 in MPV1) show relatively high identity across most Group 1 viruses (Fig. 2, Supplementary Table S2). Interestingly, some hypothetical proteins (Marpi_0299, Marpi_0338 in MPV1) are also shared among several of the Group-1 provirus genomes, suggesting these genes provide important functions for these viruses.

Genomic characterization of Group-2 proviruses: The sequences and genome organization of the three likely complete Group-2 proviruses differs considerably from that of Group-1 (Fig. 1, Fig. 3). In addition to the Group-2 provirus genomes found in *Petrotoga* sp. P8T1HF07, *Petrotoga olearia* and *Defluviitoga tunesiensis,* we identified partial proviruses in the genomes of *Petrotoga mobilis* SJ95 and *Petrotoga* sp. 9T1HF07 (Supplementary Fig. S4).

The classification based on the head-neck-tail module [30] assigned them to *Siphoviridae* of Type1 - Cluster 2. All hosts of previously described members of this virus lineage belong to Firmicutes. In agreement with this, similarity searches in Genbank revealed that these proviruses show remarkably high similarity to proviruses found in three Firmicutes genomes (Fig. 3); *Lacticigenium_naphta*e (LNV1), *Geosporobacter ferrireducens* (GFV1) and *Mahella australiensis* (MAV1).

The VIRIDIC analysis of the three full-length Group-2 proviruses (P8T1HF07V1, DTV1 and POV1), suggested they represent three new viral species (Supplementary Table S2). Interestingly, two proviruses-pairs from different phyla, P8T1HF07V1-GFV1, and DTV1-MAV1, have intergenomic similarity scores of 66.7 % and 56.7 %,

respectively, suggesting they may be assigned to the same genus (Supplementary Table S2).

The 5' module of these proviruses, encoding genes involved in virus replication and transcription, shows high level of diversity across the three Thermotogota viruses analyzed and contains several morons (i.e. proviral genes that do not have a direct viral function [39, 40]). This module is followed by highly conserved packaging, morphogenesis and lysis modules. The lysogeny module is located in the 3' end of the provirus, and also contains variable morons. The observation of many morons in Group-2 suggests that these proviruses might affect host physiology and metabolism, as well as facilitating lateral transfer of these genes between hosts. However, it should be noted that some of these morons, particularly at the 5'end, which was hard to define, might not be part of the provirus. Sequencing virus DNA isolated from capsids will resolve this issue.

The site-specific serine recombinases carried by the Group-2 proviruses are homologous, although distantly related (> 25% identity), to the ones found in the majority of the Group-1 proviruses (Fig. 4). We identified a total of 28 additional protein families shared between the two groups (Table 2, Supplementary Table S2), among these are the two methylases and the regulatory proteins discussed above that are widely distributed among the Group-1 viruses. This suggests that these distinct virus-lineages are sampling genes from a common pool.

Distribution and broad host range of Thermotogota Group-1 (pro)viruses

The Group-1 proviruses are widespread across the Thermotogota phylum (Supplementary Fig. 1B). In addition to the identified proviruses, several Thermotogota genomes with no

provirus carry close homologs of Group-1 provirus genes, e.g. the serine recombinase discussed above, suggesting a history of infection of a Group-1 virus (Fig. 4). Blast searches of hallmark virus genes [41] against a database of 59 Thermotogota genomes, with no provirus identified, identified homologs (> 50% identity and 60% coverage, see below) of typical virus genes in 50 genomes. Among these were a terminase homolog in the genome of *Thermotoga* sp. RQ7 and the recombinase in the genome of *Thermotoga* sp. SG1 (Fig. 4, Supplementary Fig. S2), suggesting Group-1 viruses also have infected bacteria from the genus *Thermotoga* where no provirus have been identified yet. Several *Thermosipho*, *Marinitoga* and *Mesotoga* genomes with no provirus also carry homologs of this recombinase (Fig. 4).

The relatively high similarity among the Group-1 proviral genomes (Fig. 2) suggest that they display broad host ranges, as observed for MPV1 from *M. piezophila*, which was shown to be able infect and transfer a plasmid to a *Thermosipho* isolate [15]. To explore the host range of these predicted or active temperate viruses, we mapped CRISPR spacer sequences from 90 Thermotogota genomes to the Group-1 genomes (Table 1). This revealed that 12 of the provirus genomes have one or more matches to CRISPR spacers in Thermotogota genomes. Three proviruses had no matches; MLV1 from *Mesoaciditoga*, the incomplete MTOLDCV1 from *Mesotoga*, GPV1 from *Geotoga*. Two viruses had matches to spacers from its host genome: MCV2, TAV1. Most of the viruses had matches to CRISPRs from genomes from the same genus as the host genome. Interestingly, five viruses had matches to spacers from a genus different than the host, supporting the notion that the Group-1 viruses have broad host ranges (Table 1). TAV1 had the most of such matches with 43 spacers from *Pseudothermotoga elfii*, *Thermosipho*

spp. and *Thermotoga* spp. mapping to the genome, suggesting that this virus has a particularly broad host range.

Anecdotal evidence also supports a broad host range of TAV1. When the sample from the Hibernia oil reservoir containing TAV1 and its host, *T. africanus* H17ap60334, was processed in the lab, we observed a mass death of *Thermotoga* isolates, originating from Troll oil reservoir in the North Sea, that were in the process of being transferred to fresh media. When sequencing the genomes of the surviving *Thermotoga* isolates [4], all survivors had at least one CRISPR spacer matching the TAV1 genome (Supplementary Fig. S5). Interestingly, these spacers are located in the middle of the CRISPR array, indicating that the spacers are not new acquisitions [42], but that the surviving isolates have encountered a very similar virus, probably while inhabiting the Troll oil reservoir. These observations suggest that TAV1 can infect both *Thermotoga* and *Thermosipho* strains. It also suggests that the Troll and the Hibernia oil reservoir harbor closely related virus populations. Similar predatory stress in geographically and geologically remote subsurface environments have been observed for *Methanohalophilus* genomes originating from reservoirs from USA and Russia [43].

Close relationship of Thermotogota Group-2 proviruses and *Firmicutes*

For the Group-2 proviruses, highly similar (65-100% amino acid identity for individual proteins) provirus were detected in three Firmicutes genomes; *Geosporobacter ferrireducens* IRF9 (NZ_CP017269), *Lacticigenium naphtae* DSM 19658 and *Mahella australiensis* 50-1 BON (NC_015520). When comparing the Thermotogota and Firmicutes proviruses it was apparent that the levels of similarity does not correspond to

phylogeny of the host (Fig. 3). Genome alignments revealed that the provirus in *Petrotoga* sp. P8T1HF07 shows highest similarity to the provirus in *G. ferrireducens* with 97% DNA identity over 11,229bp and 8,389bp aligned by MAUVE [24] covering structural proteins and the packing machinery. In comparison, for the same regions, P8T1HF07V1 show 53-75% DNA identity to the proviruses from *P. olearia* and *D. tunesiensis*. Similarly, the provirus in *D. tunesiensis* is most similar to the one from *M. australiensis* (95% and 97% identity in the same regions). These patterns suggest that these viruses are frequently transferred between cells belonging to different phyla.

In contrast to the Group-1 proviruses, no Thermotogota CRISPR spacers mapped to the Group-2 proviruses, suggesting that the Group-2 proviruses either have a more restricted host range within the Thermotogota or a more recent introduction to this phylum. With the exception of *D. tunesiensis*, the genomes carrying Group-2 proviruses originated from hydrocarbon impacted environments (Supplementary Table S1). The three Firmicutes isolates also originated from hydrocarbon impacted sites (*M. australiensis*; Riverslea oil field, Bowen-Surat basin, Australia [44], *L. naphtae*; crude oil from oil-water well in Akita, Japan [45] or oil contaminated soil (*G. ferrireducens*; Daejeon, South Korea [46]). These putative temperate viruses therefore appear to be particularly abundant in oil reservoirs and hydrocarbon impacted sites.

Importance of Thermotogota viruses in lateral gene transfer.

The Group-2 proviruses highlights the close connection between Thermotogota and Firmicutes viruses. However, the phylogenetic analysis of Group-1 provirus ORFs also demonstrates a close connection to Firmicutes or viruses that infect Firmicutes (Fig. 4,

Supplementary Fig. S2, Supplementary Fig. S3). Blast searches of all Group-1 provirus genes against nr supported this, and Firmicutes is the top scoring match of 15% of the virus ORFs outside their genus (8-30% for individual Group-1 proviruses, Supplementary Table S2). This is in agreement with previous analyses reporting large amounts of gene transfer between the Thermotogota and Firmicutes e.g. reference [7, 8].

In order to further assess the importance of these viruses in LGT, we screened genomes with no provirus for the presence of virus genes (Fig. 6, Supplementary Table S2). The Group-2 viruses had the most Blast matches, with 76 Group-2 proteins having a total of 613 matches, while 31 Group-1 proteins had a total of 262 matches. This is likely due to the Group-2 proviruses having more morons. Interestingly, Group-2 genes were abundant in *Pseudothermotoga* and *Fervidobacterium*, where no proviruses have been detected yet.

Phylogenetic analyses revealed that some of the genes are of Thermotogota origin, while for other genes, such as a cadmium transporter found in proviruses DTV1 and POV1, the virus homolog was distantly related to the homologs in most Thermotogota genomes and was likely acquired from Firmicutes (Supplementary Fig. S6). The phylogeny of the cadmium transporter also revealed that the virus gene has been transferred into *Fervidobacterium* and *Pseudothermotoga* genomes, supporting the notion that Group-2 viruses have interacted with these genera. For the Group-2 proviruses, the highest number of matches was to a moron in DTV1 annotated as Queuine tRNA-ribosyltransferase of Thermotogota origin (Supplementary Table 2S). The Group-1 gene with the highest number of matches was a moron in GPV1 from *G. petraea* annotated as an aldo/keto reductase, and also this gene appears to be of Firmicutes origin. These

analyses suggest that both Group-1 and Group-2 viruses are transferring genes both within the Thermotogota, but also between Thermotogota and Firmicutes.

Induction of TAV1 from *T. africanus* H17ap60333

Induction assays were performed on three of the putatively lysogenised *Thermotogota* isolated from petroleum reservoirs: one from Group-1 (*T. africanus* H17ap6033) and two from Group-2 (*Petrotoga* sp. P8T1HF07 and *P. olearia*). We successfully induced the provirus in *T. africanus* H17ap6033 using mitomycin C [15, 16]. TAV1 was shown to produce viral particles with a polyhedral head of ~50 nm in diameter and a flexible non-contractile tail of ~160 nm in length and ~10 nm in width (Fig. 5a). Based on tail morphology, TAV1 was classified to the order *Caudovirales* and the family *Siphoviridae*, confirming the classification using bioinformatic tools. This is similar to the three previously characterized temperate *Marinitoga* viruses, whose virion tails were just slightly longer [15, 16]. TAV1 induction produced low amounts of virions upon mitomycin C induction (data not shown). In addition to viral particles, we regularly observed membrane vesicles or toga fragments released from *T. africanus* H17ap6033 (Fig. 5b).

We also attempted, but were not successful in, inducing the provirus in *Petrotoga* sp. P8T1HF07 and *Petrotoga olearia*, using mitomycin C. However, as observed for *T. africanus* H17ap6033, large numbers of membrane vesicles of various sizes and shapes (20 - 100nm) were produced by the cells (Fig. 5c). For *Petrotoga* sp. P8T1HF07, we observed vesicles of similar round shape, connected together and extended as long chains by hooking on to the flagella, like a “pearl necklace”, while free vesicles showed more

diversity in size and shape (Fig. 5c). Some “sunflower-like” structures were also observed inside the cell (Fig. 5d). It is unknown if the provirus or environmental stressors influence the production of these vesicles and structures, or if they are produced spontaneously.

Thus to date, only Group-1 proviruses from *Marinitoga* [15, 16] and *Thermosipho* have been induced leading to the production of viral particles. The fact that we have not been able to induce the Group-2 proviruses tested, could suggest they are inactive in Thermotogota genomes or that they are in the process of domestication or deletion from the genomes. Alternatively, we may not have provided the optimal conditions for induction.

Identification of a potential receptor of the Group-2 viruses

Several types of structures on the surface of bacteria can act as virus receptors, e.g. membrane proteins, flagella, pili, carbohydrate moieties on the cell surface [47]. Most siphoviruses of Gram-negative bacteria, and some of Gram-positive bacteria, use proteinaceous receptors for adsorption [48][49]. If the Group-2 viruses uses the same protein receptor to attach to both Thermotogota and Firmicutes cells, the large phylogenetic distance between these hosts offer an opportunity to identify possible membrane protein receptors bioinformatically. It should be noted that this approach only identifies possible protein receptors, while potential shared carbohydrate receptors will not be identified. Four predicted membrane proteins with transmembrane helices where identified in all the genomes carrying a Group-2 provirus (Supplementary Fig. S7). One of these was the virus holin gene, leaving three receptors candidates; a ComEA family

DNA-binding protein, an oxaloacetate decarboxylase beta subunit and an ABC transporter ATP-binding protein. Phylogenetic analyses revealed that the ComEA homolog and the oxaloacetate decarboxylase are widely distributed among the Thermotogota (Supplementary Fig. S7). In contrast, the ABC transporter is restricted to *Petrotoga* and *Defluviitoga* (Supplementary Fig. S7), which could explain why the Group-2 proviruses are only observed in these Thermotogota genera. Inspection of the alignment of the proteins, revealed that these proteins show particularly high similarity in the last C-terminal domain which is facing the exterior of the cell and could serve as a virus target (Supplementary Fig. S8). Although experiments are needed to explore if any of these proteins functions as receptor for these viruses, we suggest that the ABC-transporter protein is the strongest candidate for a Group-2 virus receptor.

ABC-transporter proteins are, to our knowledge, not commonly identified as bacterioviral receptors. However, membrane proteins with a variety of functions have been shown to act as receptors [47]. Moreover, *Lactoccocus* viruses from the siphoviral c2 group have been shown to use membrane proteins Pip or YjaE as secondary receptors [47][50], which both show similarity to ABC2 transporter domains.

A role of membrane transport proteins and viruses in connecting Thermotogota and Firmicutes?

ABC-transporters are among the most frequently transferred genes both within the Thermotogota and also between Thermotogota and other phyla, particularly Firmicutes [51][7, 8]. The possibility that ABC transporters, and other transporters, can function as virus receptors in the Thermotogota is intriguing, especially since transporters also are

among the morons discussed above. This suggests that acquiring a new transporter, perhaps via a virus infection, might result in the cell not only acquiring a new function but also becoming susceptible to a new virus. This virus might carry another transporter gene, which can introduce yet another virus, resulting in a ratchet like process. Using transporters as receptors will therefore not only provide the virus with the wide host range inferred for both Thermotogota virus groups but could also make viruses the vehicles on the highways of gene sharing observed between the Thermotogota and Firmicutes. Membrane transport protein encoding genes, including ABC transporters, have been observed in several other viruses [52], thus this process could also operate on other genomes. This is contrary to a role commonly assigned to morons where they often confer resistance to infections by other viruses [40]. Further studies and experiments are needed to investigate if such ratchet processes are indeed occurring in natural systems. However, regardless of the functions of the morons in the *Thermotogota* proviruses, the observation of shared viruses among organisms from different phyla demonstrates that viruses are key actors in the biosphere acting as engines of evolution.

Acknowledgements

This work is supported by a Norwegian research council award (project no. 180444/V40) to C.L.N, and by the Sino-French LIA/PRC 1211 MicrobSea for J.L. Strains were obtained from the Université de Bretagne Occidentale Culture Collection (UBOCC, Plouzané, France, www.univ-brest.fr/ubocc). We thank Emily St John and Stéphane l'Haridon who provided the *Marinitoga lauensis* proviral sequence.

Competing Interests:

The authors declare no conflict of interest.

References

1. Pollo SMJ, Zhaxybayeva O, Nesbø CL. Insights into thermoadaptation and the evolution of mesophily from the bacterial phylum *Thermotogae*. *Can J Microbiol* 2015; **61**: 655–670.
2. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol* 2016; **1**: 16048.
3. Bhandari V, Gupta RS. The Phylum Thermotogae. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F (eds). *The Prokaryotes*. 2014. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 989–1015.

4. Nesbø CL, S Swithers K, Dahle H, Haverkamp TH, Birkeland N-K, Sokolova T, et al. Evidence for extensive gene flow and Thermotoga subpopulations in subsurface and marine environments. *ISME J* 2015; **9**: 1532–1542.

5. Foght JM, Gieg LM, Siddique T. The microbiology of oil sands tailings: past, present, future. *FEMS Microbiol Ecol* 2017; **93**.

6. Nesbø CL, Charchuk R, Pollo SMJ, Budwill K, Kublanov IV, Haverkamp THA, et al. Genomic analysis of the mesophilic Thermotogae genus *Mesotoga* reveals phylogeographic structure and genomic determinants of its distinct metabolism: Comparative genomic analysis of *Mesotoga*. *Environ Microbiol* 2019; **21**: 456–470.

7. Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, et al. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. *Nature* 1999; **399**: 323–329.

8. Zhaxybayeva O, Swithers KS, Lapierre P, Fournier GP, Bickhart DM, DeBoy RT, et al. On the chimeric nature, thermophilic origin, and phylogenetic placement of the Thermotogales. *Proc Natl Acad Sci* 2009; **106**: 5865–5870.

9. Nesbø CL, Dlutek M, Doolittle WF. Recombination in Thermotoga: Implications for Species Concepts and Biogeography. *Genetics* 2006; **172**: 759–769.

10. Magnabosco C, Biddle JF, Cockell CS, Jungbluth SP, Twing KI. Biogeography, Ecology, and Evolution of Deep Life. In: Orcutt BN, Daniel I, Dasgupta R (eds). *Deep Carbon*, 1st ed. 2019. Cambridge University Press, pp 524–555.

11. Anderson RE, Brazelton WJ, Baross JA. The Deep Viriosphere: Assessing the Viral Impact on Microbial Community Dynamics in the Deep Subsurface. *Rev Mineral Geochem* 2013; **75**: 649–675.

12. Cobián Güemes AG, Youle M, Cantú VA, Felts B, Nulton J, Rohwer F. Viruses as Winners in the Game of Life. *Annu Rev Virol* 2016; **3**: 197–214.

13. Engelhardt T, Kallmeyer J, Cypionka H, Engelen B. High virus-to-cell ratios indicate ongoing production of viruses in deep subsurface sediments. *ISME J* 2014; **8**: 1503–1509.

14. Walsh EA, Kirkpatrick JB, Pockalny R, Sauvage J, Spivack AJ, Murray RW, et al. Relationship of Bacterial Richness to Organic Degradation Rate and Sediment Age in Subseafloor Sediment. *Appl Environ Microbiol* 2016; **82**: 4994–4999.

15. Lossouarn J, Nesbø CL, Mercier C, Zhaxybayeva O, Johnson MS, Charchuck R, et al. 'Ménage à trois': a selfish genetic element uses a virus to propagate within Thermotogales. *Environ Microbiol* 2015; **17**: 3278–3288.

16. Mercier C, Lossouarn J, Nesbø CL, Haverkamp THA, Baudoux AC, Jebbar M, et al. Two viruses, MCV1 and MCV2, which infect Marinitoga bacteria isolated from deep-sea hydrothermal vents: functional and genomic analysis. *Environ Microbiol* 2018; **20**: 577–587.

17. Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Pillay M, et al. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res* 2014; **42**: D568–D573.

18. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 2000; **16,**: 276–277.

19. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics* 2008; **24**: 863–865.

20. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: A Fast Phage Search Tool. *Nucleic Acids Res* 2011; **39**: W347–W352.

21. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* 2012; **40**: e126–e126.

22. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016; **44**: D733–D745.

23. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009; **10**: 421.

24. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* 2010; **5**: e11147.

25. Guy L, Roat Kultima J, Andersson SGE. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* 2010; **26**: 2334–2335.

26. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. *Genome Res* 2009; **19**: 1639–1645.

27. Zimmermann L, Stephens A, Nam S-Z, Rau D, Kübler J, Lozajic M, et al. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* 2018; **430**: 2237–2243.

28. Bolduc B, Jang HB, Doulcier G, You Z-Q, Roux S, Sullivan MB. vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect *Archaea* and *Bacteria*. *PeerJ* 2017; **5**: e3243.

29. Bin Jang H, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, et al. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* 2019; **37**: 632–639.

30. Lopes A, Tavares P, Petit M-A, Guérois R, Zinn-Justin S. Automated classification of tailed bacteriophages according to their neck organization. *BMC Genomics* 2014; **15**: 1027.

31. Moraru C, Varsani A, Kropinski AM. VIRIDIC – a novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. 2020. Microbiology.

32. Nishimura Y, Yoshida T, Kuronishi M, Uehara H, Ogata H, Goto S. ViPTree: the viral proteomic tree server. *Bioinformatics* 2017; **33**: 2379–2380.

33. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; **30**: 1312–1313.

34. Geslin C, Le Romancer M, Erauso G, Gaillard M, Perrot G, Prieur D. PAV1, the First Virus-Like Particle Isolated from a Hyperthermophilic Euryarchaeote, “Pyrococcus abyssi”. *J Bacteriol* 2003; **185**: 3888–3894.

35. Haverkamp THA, Geslin C, Lossouarn J, Podosokorskaya OA, Kublanov I, Nesbø CL. Thermosipho spp. immune system differences affect variation in genome size and geographical distributions. *Genome Biol Evol* 2018.

36. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Dempsey DM, et al. Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020). *Arch Virol* 2020; s00705-020-04752–x.

37. Adriaenssens E, Brister JR. How to Name and Classify Your Phage: An Informal Guide. *Viruses* 2017; **9**: 70.

38. Catalano CE. The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. *Cell Mol Life Sci CMLS* 2000; **57**: 128–148.

39. Cumby N, Davidson AR, Maxwell KL. The moron comes of age. *Bacteriophage* 2012; **2**: e23146.

40. Taylor VL, Fitzpatrick AD, Islam Z, Maxwell KL. The Diverse Impacts of Phage Morons on Bacterial Fitness and Virulence. *Adv Virus Res* 2019; **103**: 1–31.

41. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ* 2015; **3**: e985.

42. Rath D, Amlinger L, Rath A, Lundgren M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* 2015; **117**: 119–128.

43. Borton MA, Daly RA, O'Banion B, Hoyt DW, Marcus DN, Welch S, et al. Comparative genomics and physiology of the genus *Methanohalophilus* , a prevalent methanogen in hydraulically fractured shale. *Environ Microbiol* 2018; **20**: 4596–4611.

44. Bonilla Salinas M. Mahella australiensis gen. nov., sp. nov., a moderately thermophilic anaerobic bacterium isolated from an Australian oil well. *Int J Syst Evol Microbiol* 2004; **54**: 2169–2173.

45. Iino T, Suzuki K -i., Harayama S. Lacticigenium naphtae gen. nov., sp. nov., a halotolerant and motile lactic acid bacterium isolated from crude oil. *Int J Syst Evol Microbiol* 2009; **59**: 775–780.

46. Hong H, Kim S-J, Min U-G, Lee Y-J, Kim S-G, Jung M-Y, et al. Geosporobacter ferrireducens sp. nov., an anaerobic iron-reducing bacterium isolated from an oil-contaminated site. *Antonie Van Leeuwenhoek* 2015; **107**: 971–977.

47. Stone E, Campbell K, Grant I, McAuliffe O. Understanding and Exploiting Phage–Host Interactions. *Viruses* 2019; **11**: 567.

48. Zhang Z, Yu F, Zou Y, Qiu Y, Wu A, Jiang T, et al. Phage protein receptors have multiple interaction partners and high expressions. *Bioinformatics* 2020; **36**: 2975–2979.

49. Bertozzi Silva J, Storms Z, Sauvageau D. Host receptors for bacteriophage adsorption. *FEMS Microbiol Lett* 2016; **363**: fnw002.

50. Millen AM, Romero DA. Genetic determinants of lactococcal C2viruses for host infection and their role in phage evolution. *J Gen Virol* 2016; **97**: 1998–2007.

51. Nesbø CL, Nelson KE, Doolittle WF. Suppressive Subtractive Hybridization Detects Extensive Genomic Diversity in Thermotoga maritima. *J Bacteriol* 2002; **184**: 4475–4488.

52. Greiner T, Moroni A, Van Etten J, Thiel G. Genes for Membrane Transport Proteins: Not So Rare in Viruses. *Viruses* 2018; **10**: 456.

53. L'Haridon S, Gouhier L, John ESt, Reysenbach A-L. Marinitoga lauensis sp. nov., a novel deep-sea hydrothermal vent thermophilic anaerobic heterotroph with a prophage. *Syst Appl Microbiol* 2019; **42**: 343–347.

54. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019; **47**: W256–W259.

55. Lee MD. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* 2019; **35**: 4162–4164.

Table 1. CRISPR matches to provirus genomes. Proviruses that have been induced and shown to produce virus particles are indicated by a *. Spacers were mapped to virus sequences in Geneious v.10 allowing up to 10% mismatches. Numbers of spacers from each genome mapping to the provirus is given in parentheses. Matches to spacers from the host genome are noted as a self-match.

provirus name	CRISPR spacer match from (number of spacers)
KPV1	*Kosmotoga olearia* (1)
MHV1	*Pseudothermotoga elfii* NBRC107921 (1), *Marinitoga* sp. 1154 (1)
MCV1*	*Marinitoga* sp. 1154 (2)
MCV2*	*Marinitoga* sp. 1155 (1) self-match, *Marinitoga* sp. 1154 (2)
MPV1*	*Marinitoga* sp. 1137 (1)
M1135	*Marinitoga* sp. 1155 (1)
M1137V1	*Pseudothermotoga elfii* NBRC107921 (1), *Marinitoga* sp. 1135 (2)
M1137V2	*M. piezophila* (1)
M1138	*T. africanus* TCF52B (1), *T. melanesiensis* (1), *Pseudothermotoga elfii* NBRC107921 (1), *Marinitoga* sp. 1137 (1)
TAV1*	*T. africanus* H17ap60334 (3) self -match, *T. africanus* TCF52B (2), *T. africanus* Ob7 (1), *T. melanesiensis* (2), *Thermotoga maritima* 2812B (1), *Thermotoga* sp. EMP (1), *Thermotoga* sp. XYL54 (3), *Thermotoga* sp. CELL2 (3), *Thermotoga* sp. TBGT1766 (3), *Thermotoga* sp. TBGT1765 (4), *Thermotoga* sp. A7A (1), *Pseudothermotoga elfii lettingae* (4), *Pseudothermotoga elfii* NBRC107921 (9), *Pseudothermotoga elfii* DSM9442 (2)

TMV1	*T. melanesiensis* (1) self-match, *T. africanus* TCF52B (1), *T. africanus* Ob7 (1), *Pseudothermotoga elfii lettingae* (2), *Thermotoga* sp. TBGT1765 (1), *Thermotoga* sp. Mc24 (1), *Pseudothermotoga elfii* NBRC107921 (1)
T1074V1	*Thermosipho affectus* Bl1070 (1), *Thermosipho affectus* 1223 (3), *Thermosipho affectus* Bl1063 (1)

Table 2. Homologous genes detected among Group-1 and Group-2 provirus proteins using blastp and expected value cut-off at $1e^{-01}$. For Recombinase homologs see Fig. 5.

Annotation	**Group 1 homologs**	**Group 2 homologs**
Aldo/keto reductase	BLS00_RS05685 (GPV1)	X929_01195 (POV1)
ATP-binding protein	BLS00_RS05500 (GPV1), IX53_RS00630 (KPV1)	DTL3_1118 (DTV1)
Capsid protein	LN42_00565 (M1137V1), LH53_05775 (MTOLDC1)	NAAC61_01950 (P8T1HF07V1), X929_01335 (POV1), DTL3_1133 (DTV1)
Cro/Cl family transcriptional regulator	EK18DRAFT_00606 (MLV1), H17ap60334_04807 (TAV1), BUA62_RS02415 (MHV1), LN42_00465 (M1137V1)	NAAC61_02080 (P8T1HF07V1)
DNA methylase	BLS00_RS05440 (GPV1), BUA62_RS02350 (MHV1),	NAAC61_02100 (P8T1HF07V1)

	EK18DRAFT_00547 (MLV1)	
DNA methylase	Tmel_1480 (TMV1), Marpi_0296 (MPV1), XO11_00615 (M1138V1), EK18DRAFT_00605 (MLV1), LH53_05700 (MTOLDCV1)	DTL3_1116, DTL3_1121, DTL3_1122 (DTV1), NAAC61_01990 (P8T1HF07V1), X929_01285 (POV1)
Endonuclease-like protein	Tmel_1473 (TMV1)	NAAC61_02075 (P8T1HF07V1)
HNH endonuclease	EK18DRAFT_00580 (MLV1)	NAAC61_02005 (P8T1HF07V1), X929_01255 (POV1)
Hypothetical protein/ Helicase	XO11_00565 (M1138V1), XO10_00565 (M1135V1)	X929_01210 (POV1)
Hypothetical protein	UF08_38 (MCV1)	X925_04785 (P9T1HF07V1)
Hypothetical protein	XO08_07915 (T1074V1), LH53_05685 (MTOLDCV1)	DTL3_1144 (DTV1), NAAC61_01815* (P8T1HF07V1)
Hypothetical protein	Marpi_0334 (MPV1), LN42_00625 (M1137V1), XO11_00700 (M1138V1), H17ap60334_04967 (TAV1), Tmel_1449 (TMV1), XO08_07890 (T1074V1), LH53_05725 (MTOLDCV1),	DTL3_1141 (DTV1)

	EK18DRAFT_00564 (MAV1)	
Hypothetical protein	XO08_07730 (T1074V1)	DTL3_1114 (DTV1)
Hypothetical protein	LH53_05765 (MTOLDCV1)	DTL3_1134 (DTV1)
Hypothetical protein	EK18DRAFT_00608 (MAV1)	DTL3_1111 (DTV1)
Hypothetical protein	EK18DRAFT_00582 (MAV1)	X929_01400 (POV1)
Hypothetical protein	IX53_RS00750 (KPV1)	NAAC61_02025 (P8T1HF07V1)
Hypothetical protein	IX53_RS00690 (KPV1)	Pmob_0430* (PMV1)
Hypothetical protein	LN42_01810 (M1137V2)	NAAC61_02035 (P8T1HF07V1)
Hypothetical protein	Marpi_0313 (MPV1)	DTL3_1106 (DTV1)
Hypothetical protein/ transporter	Marpi_0315 (MPV1), XO10_00580 (1135V1)	DTL3_1155 (DTV1), X929_01440 (POV1), X925_04770 (P9T1HF07V1), Pmob_0433 (PMV1)
Phage minor structural protein	BUA62_RS02360 (MHV1)	X929_01385 (POV1), DTL3_1143 (DTV1)
Phage portal protein	LH53_05785 (MTOLDCV1)	NAAC61_01960 (P8T1HF07V1), X929_01325 (POV1), DTL3_1131 (DTV1)
Primase / Terminase	EK18DRAFT_00577 (MAV1), IX53_RS00695 (KPV1)	X929_01235 (POV1)
Tail protein	UF08_32 (MCV1), BUA62_RS02555 (MHV1),	X929_01375 (POV1)

	Tmel_1449 (TMV1), H17ap60334_04967 (TAV1)	
Transcription regulator	Marpi_0298 Marpi_0297 (MPV1), UF08_76 UF08_59 UF08_65 (MCV1), X274_02275 (MCV2), XO08_07670 (T1074V1), BUA62_RS02435 BUA62_RS02375 BUA62_RS02385 (MHV1), XO11_00475 (M1138), XO10_00515 XO10_00685 (M1135V1), LN42_01800 LN42_01805 LN42_01820 (M1137V2), H17ap60334_04807 (TAV1), Tmel_1478 Tmel_1479 (TMV1)	DTL3_1158 (DTV1)
Transposase	Tmel_1465 (TMV1)	DTL3_1117 (DTV1)
XRE family transcriptional regulator	XO11_00475 (M1138V1), UF08_9 (MCV1), XO10_00475 (M1135)	NAAC61_02010 (P8T1HF07V1)

*recombinase

Fig. 1. **Comparison of proviruses identified in eight different Thermotogota genera**. For genera where multiple different viruses were detected, only one virus is represented as viruses from the same genus were highly similar (Fig. 2 and 3). For this figure all tblastx matches between the viruses were drawn as links with the percent identity shown in histograms attached to each link. Blast matches were drawn in the following order; DTV1 (dark blue), GPV1 (grey), TV1 (orange), MPV1(red), KPV1 (dark green), MTOLDCV1 (light green), MLV1 (purple) and P8T1HF07V1 (light blue). Reciprocal best matches are therefore hidden in many cases (e.g. the dark blue links between DTV1 and P8T1HF07V1), however, these are still visible in the histogram connecting to these links. The figure was drawn in Circos [26].

Fig. 2. Comparison of the Thermotogota Group-1 provirus sequences. (Pro)virus name and the genus of the host it belongs to are shown on the left side. Proviruses that have been induced and shown to produce virus particles are indicated by a *. Proviruses determined to be incomplete are indicated by (in). The origin of the sequences are (accession number); GPV1 - *Geotoga petraea* WG14 (FMYV01), MPV1 - *Marinitoga piezophila* (NC_016751), MCV1 – *Marinitoga camini* 1197 (AZAY01), MCV2 - *Marinitoga camini* 1155 (AZAX01), MHV1 - *Marinitoga hydrogenitolerans* DSM 16785 (FQUI01), M1137V2 and M1137V2 - *Marinitoga* sp. 1137 (CP009471), M1135V1 - *Marinitoga* sp. 1135 (LBEZ01), M1138V1 - *Marinitoga* sp. 1138 (LBFA01), MLaV1 - *Marinitoga lauensis* (QHLX01), TMV1 - *Thermosipho melanesiensis* BI429 (CP000716), TAV1 – *Thermosipho africanus* H17ap60334 (AJIP01), T1074V1 –

Thermosipho sp. 1074 (LBEX01), KPV1 – *Kosmotoga pacifica* SLHLJ1 (CP011232), MAV1 - *Mesoaciditoga lauensis* cd-1655R (JQJI01) and MTOLDCV1 – *Mesotoga* sp. SC_TOLDC (AYSI01). We also included in this figure, MLaV1 - *Marinitoga lauensis* (QHLX01) reported in reference [53] after our bioinformatic screening was performed. The regions with significant pairwise tblastx similarity scores are connected (minimum identity: 30 %, minimum lenght : 100 nucleotides), red indicates that sequence is in the same direction while blue indicates that the similar sequences are on opposite strands. The division of the provirus genomes into functional modules is approximate, as it is based only on gene annotations. Gene functions are color-coded as follows; red: integration, yellow: transcriptional regulation, orange: DNA metabolism, green: DNA packaging and head, light blue: head to tail, dark blue: tail, beige: HNH endonuclease, pink: lysis, black: other (morons), grey: hypothetical proteins). Some gene annotations are included and abbreviated; Ser recomb: serine recombinase, LexA: LexA repressor, ParB: ParB-like nuclease, RecT: RecT family recombinase, dUTP hydrolase: deoxyuridine 5'-triphosphate nucleotidohydrolase, ssb: single stranded DNA-binding protein, tss: terminase small subunit, tls: terminase large subunit, mcp: major capsid protein, tail tape: tail tape measure protein, rRNA lsm: ribosomal RNA large subunit methyltransferase, flagella bbp: flagella basal-body protein, DnaC: DnaC replication protein, DnaD: DnaD replication protein, DNA pol sc: DNA polymerase sliding clamp, SecB: SecB protein-export protein, rep organizer :replisome organizer, RusA: RusA family crossover junction endodeoxyribonuclease, Cys peptidase: cysteine peptidase, sulfate AT: sulfate adenylyltransferase subunit 2, PAPS reductase: phosphoadenosine phosphosulfate reductase, CW hydrolase: cell wall-associated hydrolase, MazF: MazF

endoribonuclease, dsbr: DNA double-strand break repair protein, metal bp: metal-binding protein, CMP hydrolase: cytidine 5′-monophosphate hydrolase. The figure was produced using genoPlotR (Guy *et al*., 2010).

Fig. 3. Comparison of the putative Thermotogota and Group-2 provirus sequences and their Firmicutes closest relatives. Virus name and the genus the host belongs to are included. The origin of the sequences are (accession number); DTV1 - *Defluviitoga tunisiensis* DTL3 (LN824141), POV1 – *Petrotoga olearia* DSM 13574 (AZRL01), LNV1 – *Lacticigenium naphtae* DSM 19658 (AUEG01), P8T1HF07V1 - *Petrotoga* sp. 8T1HF07 (AUPG01), GSV1 - *Geosporobacter ferrireducens* IRF9 (CP017269), MAUV1 - *Mahella australiensis* 50-1 BON (NC_015520). The regions with significant pairwise tblastx similarity scores are connected (mininum identity: 30 %, minimum lenght: 100 nucleotides), red indicates that sequence is in the same direction while blue indicates that the similar sequences are on opposite strands. The division of the provirus genomes into functional modules is approximate, as it is based only on gene annotations. Selected informative ORF annotations are shown. Gene functions are color-coded: red: integration, yellow: transcriptional regulation, orange: DNA metabolism, green: DNA packaging and head, light blue: head to tail, dark blue: tail, white: HNH endonuclease, pink: lysis, black: other (morons), purple: transposase, grey: hypothetical proteins). Some gene annotations are included and abbreviated; HMT ATPase: heavy metal translocating ATPase, FMN reductase: flavine mono nucleotide reductase, HAD family phosphatase: haloacid dehalogenase superfamily of hydrolase). The figure was produced using genoPlotR (Guy *et al*., 2010).

Fig. 4. Maximum likelihood tree of recombinase genes. Sequences homologous to the recombinase from all Thermotogota viruses and the three Firmicutes viruses in Fig. 4 were identified by blastp searches against Genbank and the twenty top matches of each virus homolog were retrived. The resulting dataset was filtered keeping only unique sequences and the sequences were aligned using MAFFT as implemented in Geneious v. 10. A FastTree analysis was performed first in Genenious v. 10 and clades containing numerous almost identical Firmicute sequences were trimmed. The phylogenetic tree displayed was constructed using RAxML as implemented in Geneious v. 10 with a GAMMA WAG model and 100 bootstrap replicates. Bootstrap support > 70% is shown on branches as circles, with the size corresponding to the strength of support. (Pro)virus name is added in to the host name for proteins from the Thermotogota proviruses (e.g. ***Thermosipho africanus* TAV1**), as well as from the Firmicute proviruses in Fig. 3. Incomplete viruses are indicated by (in). Thermotogota homologs from genomes with no detected provirus are indicated by a star. Numbers attached to each sequence represent database accession numbers. Some proteins have identical sequences in more than one organisms. The protein labelled 'Bacteria inc. Petrotoga olearia POLV1 *Fervidobacterium* spp.'corresponds to accession number WP_011994748.1 and is found in *Fervidobacterium nodosum* Rt17-B1 (NC_009718.1), *Fervidobacterium pennivorans* DSM 9078 (NC_017095.1), *Fervidobacterium islandicum* (NZ_CP014334.1) and *Fervidobacterium gondwane*nse DSM 13020 (FRDJ01), *Petrotoga olearia* (PNR98053) and *Coprothermobacter proteolyticus* (PXJB01). The protein labelled 'Bacteria inc. *Mahella australiensis* MAV1, *Pseudothermotoga elfii*' corresponds to accession number

WP_013782344.1 found in *Clostridium* sp. SYSU GA15002T (NZ_CP040924.1) and *Thermoanaerobacter thermocopriae* JCM 7501 (NZ_KI912455.1), *Pseudothermotoga elfii* and MAV1 from *Mahella australiensis.* All branches without labels represent Firmicutes sequences. The tree was visualized in iTOL [54].

Fig. 5. Electron micrographs, stained with 2% uranyl acetate, of the TAV1 virus particle (panel a) and vesicles produced by *Thermosipho africanus* H17ap60334 (panel b) and *Petrotoga* sp. 8T1HF07.NaAc.6.1 (panel c, d). The TAV1 virus particle shows a typical Siphoviridae morphology (panel a). The virus particles were observed after adding 5µg/ml mitomycin C at the beginning of the exponential growth phase of *Thermosipho africanus* H17ap60334 growing at 65°C. Induction of the provirus in *Petrotoga* sp. 8T1HF07.NaAc.6.1 was attempted by adding 5µg/ml mitomycin C at the beginning of the exponential growth phase at 55°C. No virus particles were observed, however, large amounts of vesicles were observed (panel b and d), as well as sunflower-like structures (arrows in panel c). Panel b shows the vesicles attached to the flagellum.

Fig. 6. Virus genes in genomes with no provirus. Predicted proteins from proviruses were compared to a database of protein from 59 Thermotogota genomes (from eight genera) with no proviruses using blastp. This resulted in a total 875 matches for 107 provirus proteins to proteins from 50 genomes (> 50% identity and 60% coverage). The Thermotogota genera represented in this analysis is shown on the x-axis, and the number of provirus matches pr genome searched for each genus is shown on y-axis.

Supplementary Figures:

Supplementary Fig. S1. A: Network of provirus genomes calculated in VContact2. Proviruses a maximum of three nodes away from MPV1 and P8T1HF07V1 are show. The Thermotogota proviruses are colored blue, viruses from Firmicutes are green and viruses infecting other taxa are colored orange. The quality score calculated by ClusterOne for the Group-1 cluster was 0.94 and p=0.00004, the quality score for the Group-2 cluster was 0.83 and p=0.00557971. **B: Phylogenetic tree of Thermotogota genomes**. Identified proviruses are indicated next to their respective host genera. Available Thermotogota genomes in Genbank, and a list of assembly accession numbers were used as input for the GToTree pipeline, which downloads the genomes, extracts single copy gene (SCGs) sets and aligns them based on a given hmm-gene set [55]. The Bacterial hmm-set with 74 target genes was used, and the resulting alignment was imported into Geneious Prime 2020.1.2 where sites with more than 50% gaps were removed, resulting in an alignment of 11,003 amino acid positions. The phylogenetic tree was constructed using FastTree and visualized in iTOL [54].

Supplementary Fig. S2. The region of the phylogenetic tree, calculated from virus proteins by Vip-tree [32], that include the Group-1 and Group-2 Thermotogota and Firmicutes proviruses that were determined to be likely complete (indicated by red stars). Taxonomy of the most closely related viruses and their host is indicated as color bars in front of the virus name for each terminal leaf on the tree.

Supplementary Fig. S3. Maximum likelihood trees of Terminase genes. Sequences homologous to the Terminases from all Thermotogota Group 1 viruses were identified by blastp searches against Genbank retrieving the twenty top matches of each virus homolog. The resulting dataset was filtered keeping only unique sequences and the sequences were aligned using MAFFT as implemented in Geneious v. 10. A FastTree analysis was performed first in Genenious v. 10 and clades containing numerous almost identical sequences were trimmed. The phylogenetic trees displayed were constructed using RAxML as implemented in Geneious v. 10 with a GAMMA-WAG substitution model and 100 bootstrap replicates. Bootstrap support > 70% is shown on branches as circles, with the size corresponding to the strength of support. Virus name is added in bold to the host name in the tree for the Thermotogota viruses (e.g. ***Thermosipho africanus* TAV1**). Thermotogota homologs from genomes with no detected provirus are also given in bold (e.g. ***Thermotoga* sp. RQ7**). Numbers attached to each sequence represent database accession numbers. The tree was visualized in iTOL [54].

Supplementary Fig. S4. Comparison of the Thermotogota Group 2 provirus sequences including two incomplete ones. Virus name and the genus the host belongs to is indicated. Proviruses determined to be incomplete are indicated by (in). The origin of the sequences are (accession number in parenthesis); DTV1 - *Defluviitoga tunisiensis* DTL3 (LN824141), POV1 – *Petrotoga olearia* DSM 13574 (AZRL01), P8T1HF07V1 - *Petrotoga* sp. 8T1HF07 (AUPG01), PMOV1 - *Petrotoga olearia* (NZ_AZRL01), P9T1HF07V1 - *Petrotoga* sp. P9T1HF07 (AZRI01). The regions with significant pairwise BLASTN similarity scores are connected, red indicates that sequence is in the

same direction while blue indicates that the similar sequences are on opposite strands. Selected informative ORF annotations are shown. The division of the virus into functional modules is approximate, as it is based only on gene annotations. The Fig. was produced using genoPlotR (Guy *et al*., 2010).

Supplementary Fig. S5. Overview of CRISPR spacer sequences from Thermotoga isolates from the Troll oil reservoir mapped on to the TAV1 genome. Mapping and visualization was performed in Geneious v. 10 and maximum of one mismatch was allowed.

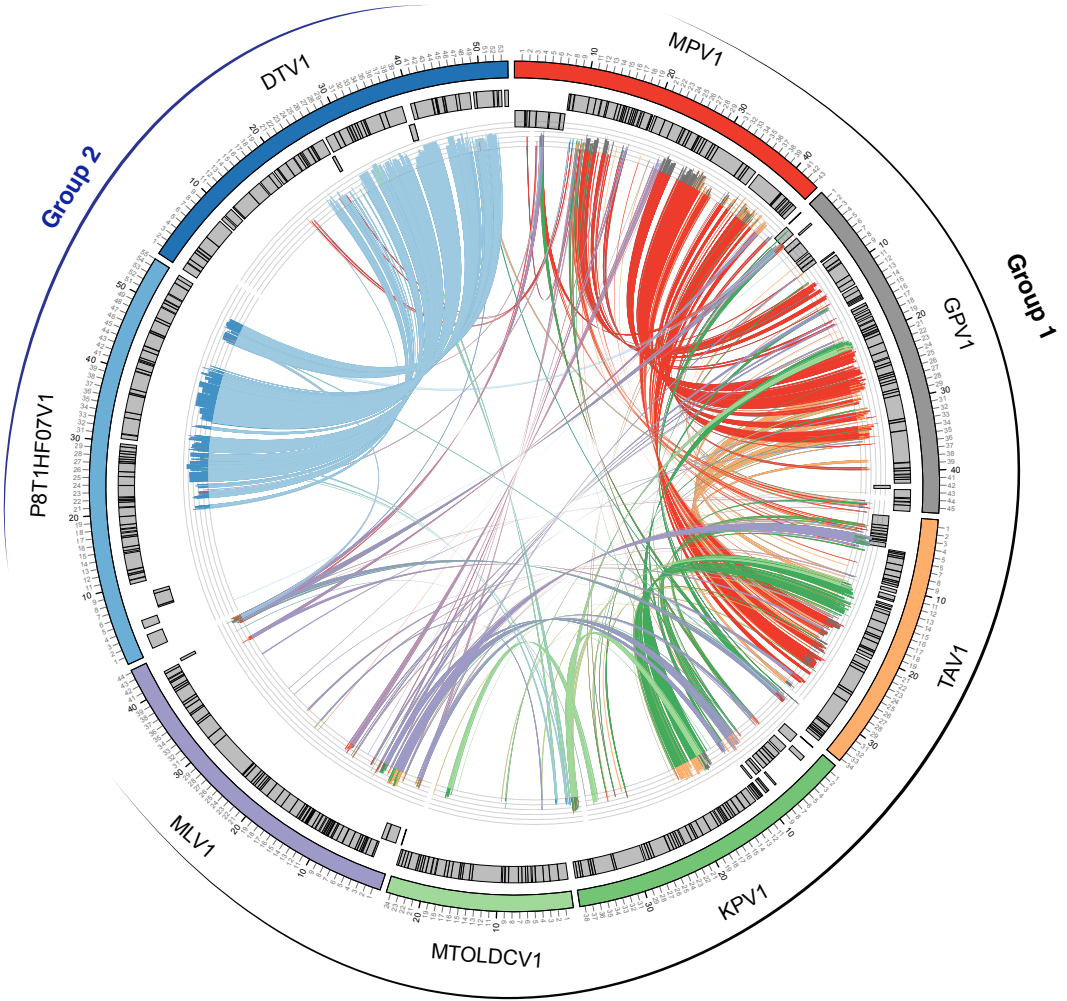
Supplementary Fig. S6. Maximum likelihood trees of the moron annotated as a cadmium transporter. Sequences homologous to the gene from all Group 2 viruses were identified by blastp searches against Genbank retrieving the twenty top matches of each virus homolog as well as searching a custom database containing genes from both Thermotogota genomes with or without a provirus. The resulting dataset was filtered keeping only unique sequences and the sequences were aligned using MAFFT as implemented in Geneious v. 10. A FastTree analysis was performed first in Genenious v. 10 and clades containing numerous almost identical sequences were trimmed. The phylogenetic trees displayed were constructed using RAxML as implemented in Geneious v. 10 with a GAMMA WAG model and 100 bootstrap replicates. Bootstrap support > 70% is shown on branches as circles, with the size corresponding to the strength of support. Virus name is added in bold to the host name in the tree for the Thermotogota viruses (e.g. ***Defluviitoga tunisiensis* DTV1**). Thermotogota homologs

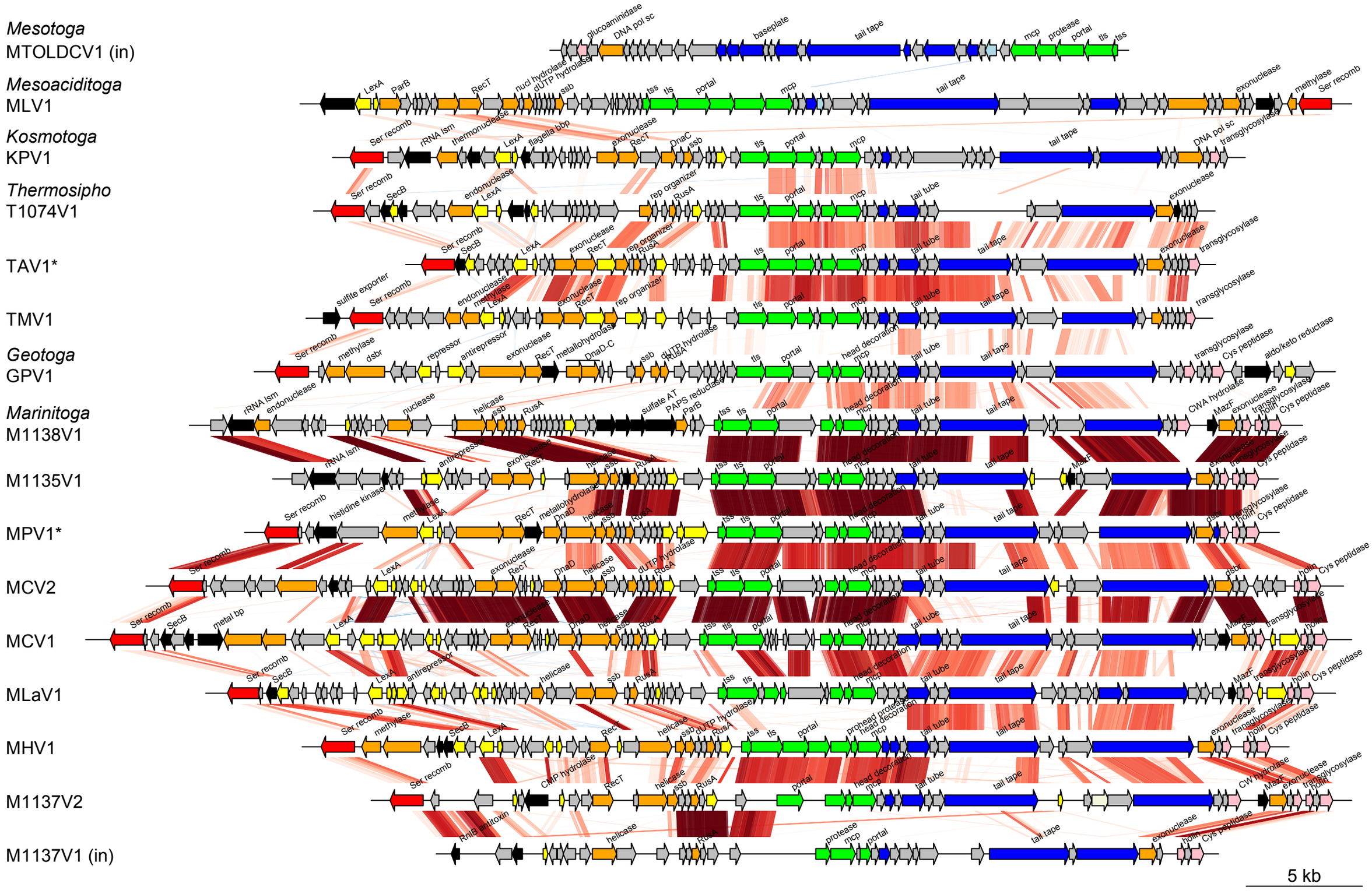
from genomes with no detected provirus are also given in bold (e.g. ***Fervidobacterium* sp. SC NGM5 G05**). Numbers attached to each sequence represent database accession numbers. The tree was visualized in iTOL [54].

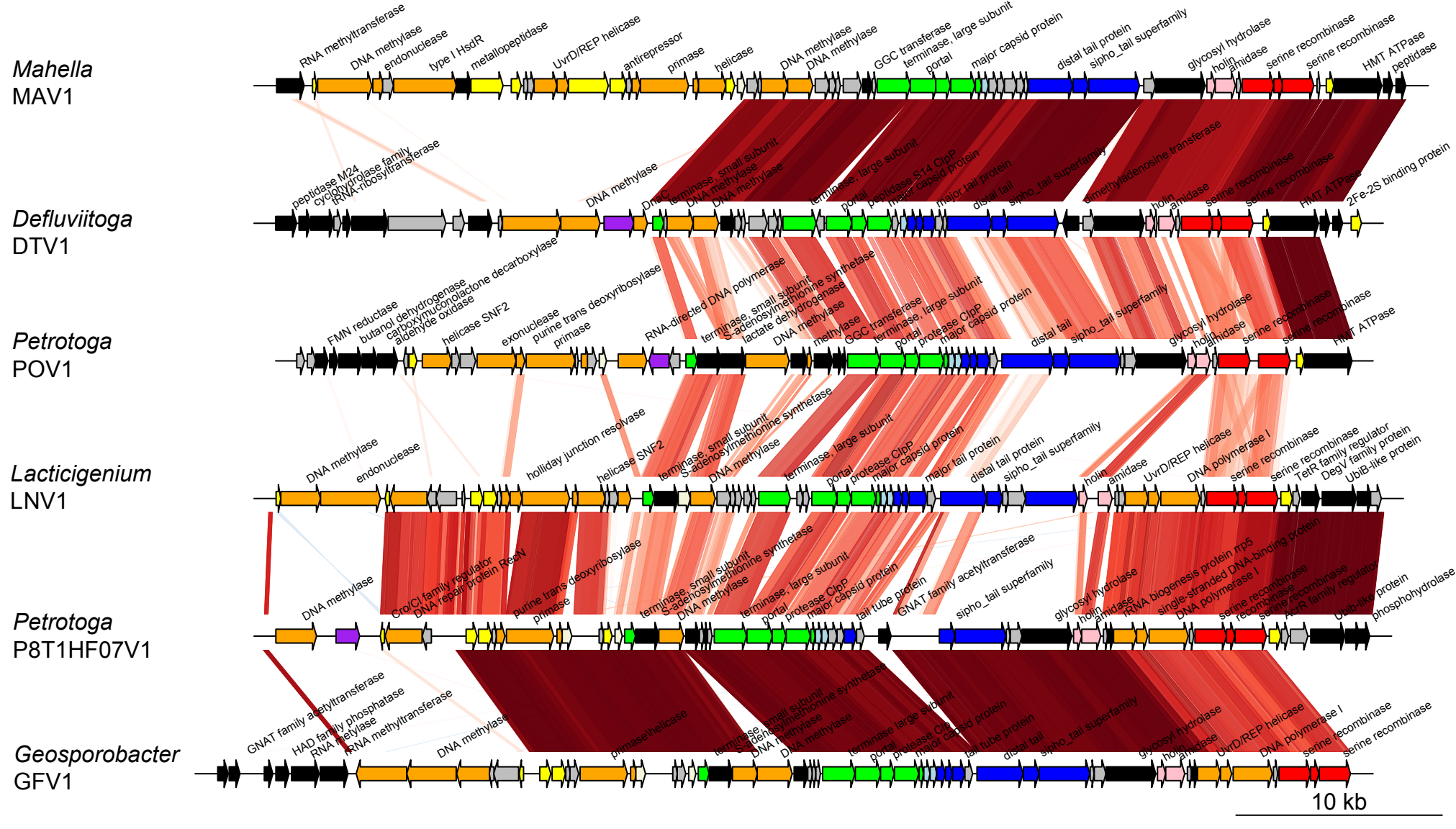
Supplementary Fig. S7. Maximum likelihood trees of potential virus receptor genes; Panel A: ComEA, Panel B: oxaloacetate decarboxylase and Panel C: ATP-binding cassette, subfamily B. Sequences homologous to the genes from the *Petrotoga*, *Defluviitoga* and the three Firmicute genomes carrying Group-2 proviruses were identified by blastp searches against Genbank, retrieving the twenty top matches of each homolog, as well as the hundred top matches from *Petrotoga* sp. 8T1HF07 to ensure as many Thermotogota homologes as possible were included. The resulting dataset was filtered keeping only unique sequences and the sequences were aligned using MAFFT as implemented in Geneious v. 10. The phylogenetic trees displayed were constructed using RAxML as implemented in Geneious v. 10 with a GAMMA WAG model and 100 bootstrap replicates. Bootstrap support > 70% is shown on branches as circles, with the size corresponding to the strength of support. The genomes with Group-2 viruses are displayed in bold and clades containing sequences from the same genus or with identical taxonomic assignment are collapsed. The trees were visualized in iTOL [54].

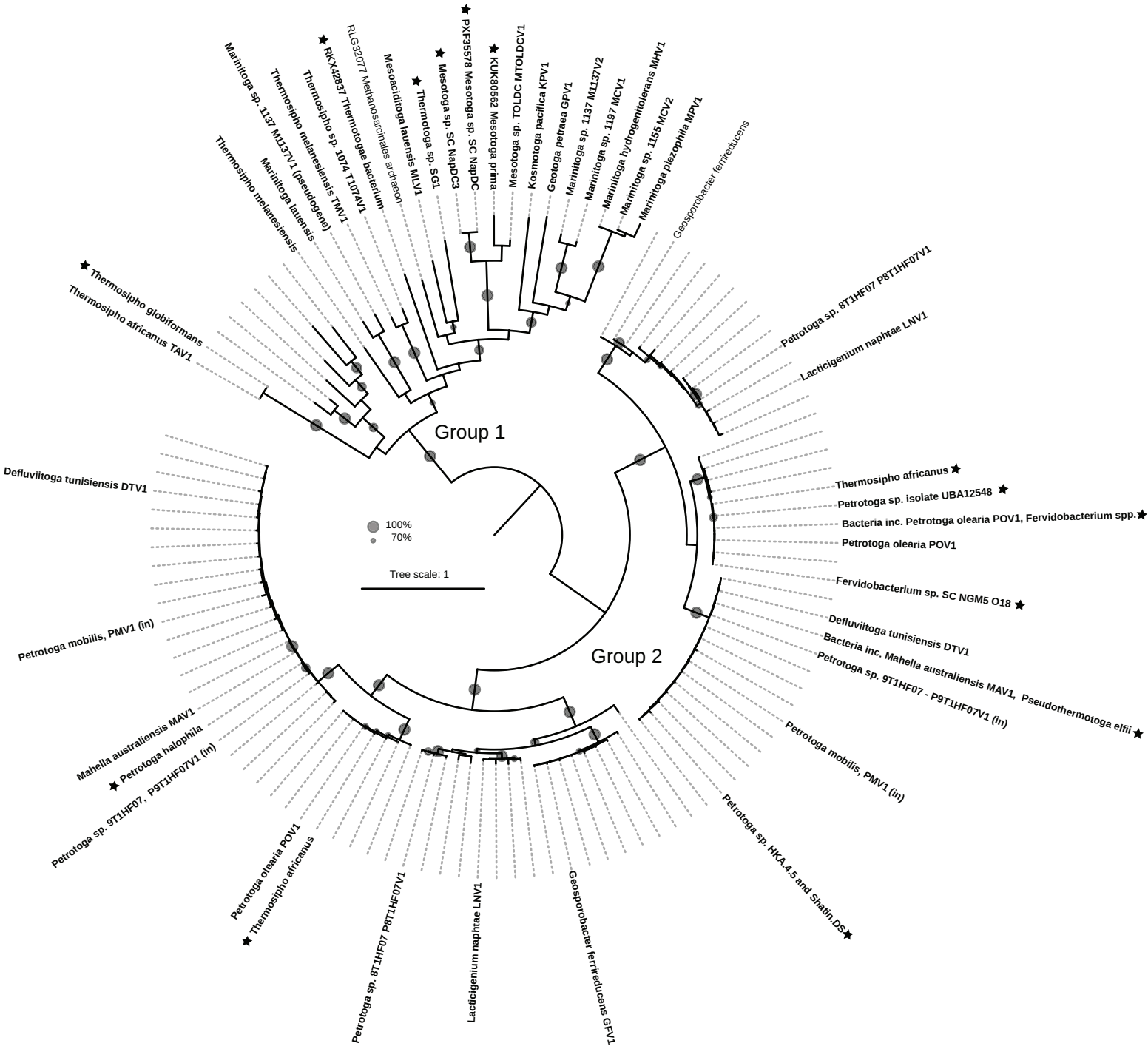
Supplementary Fig. S8. Alignment of ATP-binding cassette, subfamily B homologs. Sites identical among all the sequences are highlighted. Transmembrane regions were predicted by TMHMM v. 2 at the TMHMM Server v. 2.0

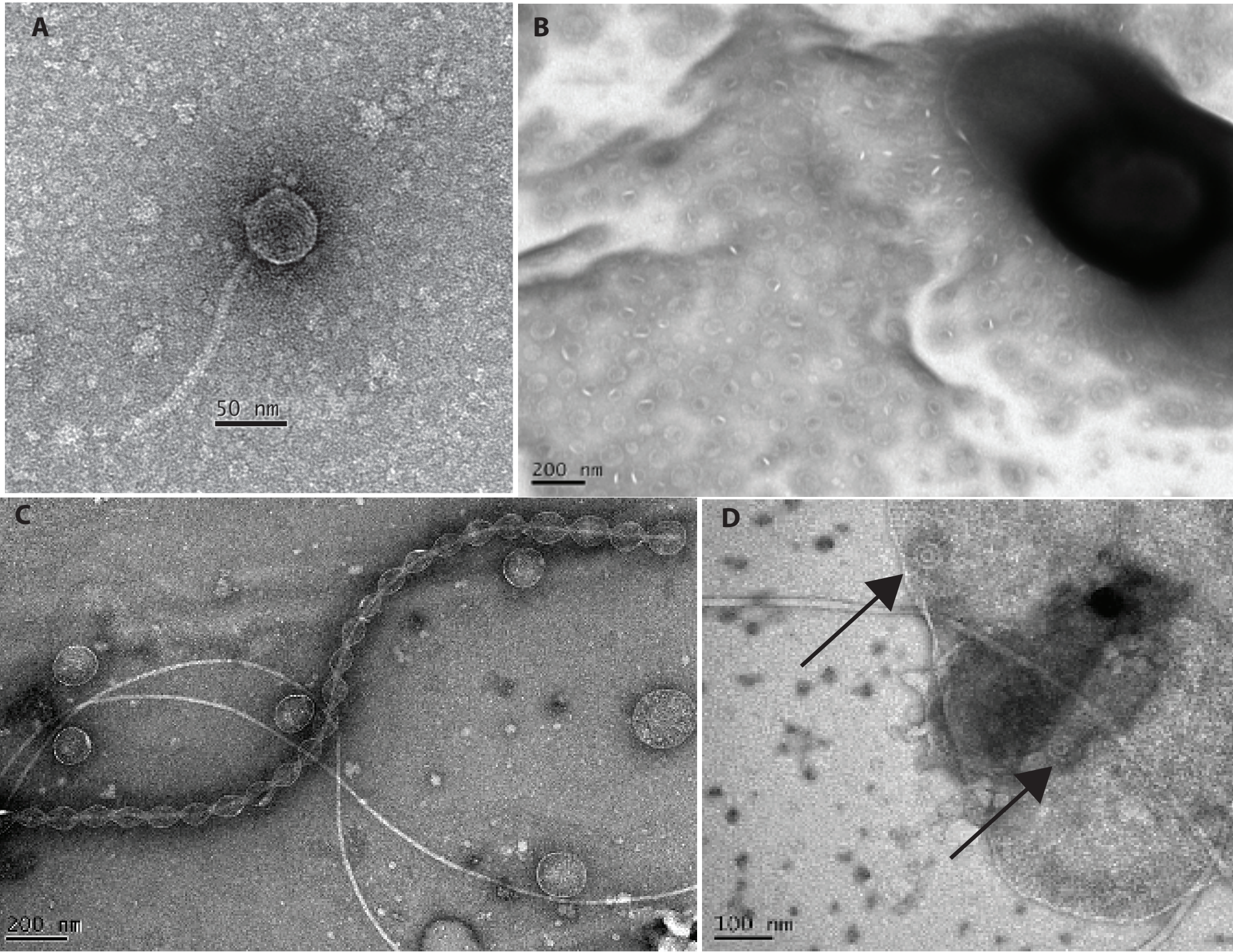
(http://www.cbs.dtu.dk/services/TMHMM-2.0/). Identical sites are highlighted with the same color and variable sites are in grey.

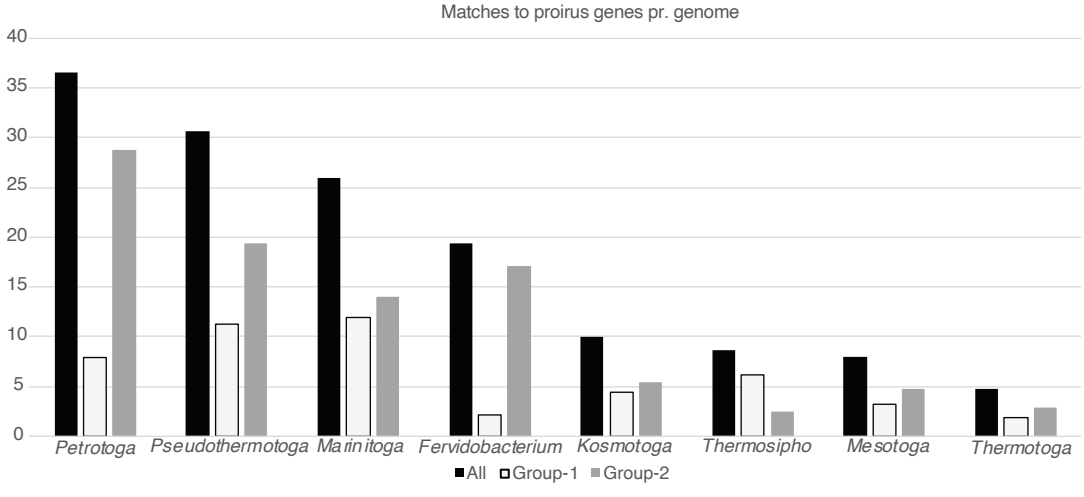

Group 2
Group 1
DTV1
MPV1
GPV1
TAV1
KPV1
MTOLDCV1
MLV1
P8T1HF07V1


Mesotoga
MTOLDCV1 (in)
Mesoaciditoga
MLV1
Kosmotoga
KPV1
Thermosipho
T1074V1
TAV1*
TMV1
Geotoga
GPV1
Marinitoga
M1138V1
M1135V1
MPV1*
MCV2
MCV1
MLaV1
MHV1
M1137V2
M1137V1 (in)
5 kb


Mahella
MAV1
RNA methyltransferase
DNA methylase
endonuclease
type I HsdR
metallopeptidase
UvrD/REP helicase
antirepressor
primase
helicase
DNA methylase
DNA methylase
GGC transferase
terminase, large subunit
portal
major capsid protein
distal tail protein
sipho_tail superfamily
glycosyl hydrolase
holin
amidase
serine recombinase
serine recombinase
HMT ATPase
peptidase
Defluviitoga
DTV1
peptidase M24
cyclohydrolase family
tRNA-ribosyltransferase
DNA methylase
DnaC
terminase, small subunit
DNA methylase
DNA methylase
terminase, large subunit
portal
peptidase S14 ClpP
major capsid protein
major tail protein
distal tail
sipho_tail superfamily
dimethylladenosine transferase
holin
amidase
serine recombinase
serine recombinase
HMT ATPase
2Fe-2S binding protein
Petrotoga
POV1
FMN reductase
butanol dehydrogenase
carboxymuconolactone decarboxylase
aldehyde oxidase
helicase SNF2
exonuclease
purine trans deoxyribosylase
primase
RNA-directed DNA polymerase
terminase, small subunit
S-adenosylmethionine synthetase
lactate dehydrogenase
DNA methylase
methylase
GGC transferase
terminase, large subunit
portal
protease ClpP
major capsid protein
distal tail
sipho_tail superfamily
glycosyl hydrolase
holin
amidase
serine recombinase
serine recombinase
HMT ATPase
Lacticigenium
LNV1
DNA methylase
endonuclease
holliday junction resolvase
helicase SNF2
terminase, small subunit
S-adenosylmethionine synthetase
DNA methylase
terminase, large subunit
portal
protease ClpP
major capsid protein
major tail protein
distal tail protein
sipho_tail superfamily
holin
amidase
UvrD/REP helicase
DNA polymerase I
serine recombinase
serine recombinase
TetR family regulator
DegV family protein
UbiB-like protein
Petrotoga
P8T1HF07V1
DNA methylase
Cro/CI family regulator
DNA repair protein RecN
purine trans deoxyribosylase
primase
terminase, small subunit
S-adenosylmethionine synthetase
DNA methylase
terminase, large subunit
portal
protease ClpP
major capsid protein
tail tube protein
GNAT family acetyltransferase
sipho_tail superfamily
glycosyl hydrolase
holin
amidase
tRNA biogenesis protein rp5
single-stranded DNA-binding protein
DNA polymerase I
serine recombinase
recombinase
serine recombinase
KcsR family regulator
Ubib-like protein
phosphohydrolase
Geosporobacter
GFV1
GNAT family acetyltransferase
HAD family phosphatase
RNA methylase
tRNA methyltransferase
DNA methylase
primase helicase
terminase, small subunit
S-adenosylmethionine synthetase
DNA methylase
DNA methylase
terminase large subunit
portal
protease ClpP
major capsid protein
tail tube protein
distal tail
sipho_tail superfamily
glycosyl hydrolase
holin
amidase
UvrD/REP helicase
DNA polymerase I
serine recombinase
serine recombinase
10 kb


Group 1
Group 2
100%
70%
Tree scale: 1
Thermosipho globiformans
Thermosipho africanus TAV1
Thermosipho melanesiensis
Marinitoga lauensis
Marinitoga sp. 1137 M1137V1 (pseudogene)
Thermosipho melanesiensis TMV1
Thermosipho sp. 1074 T1074V1
RKX42837 Thermotogae bacterium
RLG32077 Methanosarcinales archaeon
Mesaciditoga lauensis MLV1
Thermotoga sp. SG1
Mesotoga sp. SC NapDC3
PXF35578 Mesotoga sp. SC NapDC
KUK80562 Mesotoga prima
Mesotoga sp. TOLDC MTOLDCV1
Kosmotoga pacifica KPV1
Geotoga petraea GPV1
Marinitoga sp. 1137 M1137V2
Marinitoga sp. 1197 MCV1
Marinitoga hydrogenitolerans MHV1
Marinitoga sp. 1155 MCV2
Marinitoga piezophila MPV1
Geosporobacter ferrireducens
Petrotoga sp. 8T1HF07 P8T1HF07V1
Lacticigenium naphtae LNV1
Thermosipho africanus
Petrotoga sp. isolate UBA12548
Bacteria inc. Petrotoga olearia POV1, Fervidobacterium spp.
Petrotoga olearia POV1
Fervidobacterium sp. SC NGM5 O18
Defluviitoga tunisiensis DTV1
Bacteria inc. Mahella australiensis MAV1, Pseudothermotoga elfii
Petrotoga sp. 9T1HF07 - P9T1HF07V1 (in)
Petrotoga mobilis, PMV1 (in)
Petrotoga sp. HKA.4.5 and Shatin.DS
Geosporobacter ferrireducens GFV1
Lacticigenium naphtae LNV1
Petrotoga sp. 8T1HF07 P8T1HF07V1
Petrotoga olearia POV1
Thermosipho africanus
Petrotoga sp. 9T1HF07, P9T1HF07V1 (in)
Petrotoga halophila
Mahella australiensis MAV1
Petrotoga mobilis, PMV1 (in)
Defluviitoga tunisiensis DTV1


A
50 nm
B
200 nm
C
200 nm
D
100 nm


Matches to proirus genes pr. genome
40
35
30
25
20
15
10
5
0
Petrotoga
Pseudothermotoga
Marinitoga
Fervidobacterium
Kosmotoga
Thermosipho
Mesotoga
Thermotoga
All
Group-1
Group-2