Newly identified proviruses in Thermotogota suggest that viruses are the vehicles on the highways of interphylum gene sharing

Thomas H A Haverkamp, Julien Lossouarn, Olga Zhaxybayeva, Jie Lyu, Nadège Bienvenu, Claire Geslin, Camilla L Nesbø

Newly identified proviruses in Thermotogota suggest that viruses are the vehicles on the highways of interphylum gene sharing.

Thomas H. A. Haverkamp[a#], Julien Lossouarn [b], Olga Zhaxybayeva[c], Jie Lyu[d], Nadège Bienvenu[d], Claire Geslin[d] and 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] Université 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] Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON, Canada, M5S 3E5

*Corresponding Author:

nesbo@ualberta.ca

Department of Biological Sciences, CW 405 Biological Sciences Bldg. , 11455 17, Saskatchewan Drive , University of Alberta, Edmonton, Alberta, Canada, T6G 2E9

Running head: Viruses and Proviruses of Thermotogota.

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

Originality-Significance Statement:

Bacteria from the phylum Thermotogota are ubiquitous members of the very populated, but poorly understood subsurface environment. Even less understood are viruses that infect this phylum. In this study we identified 17 novel proviruses from eight different genera (*Marinitoga*, *Thermosipho*, *Kosmotoga*, *Mesotoga*, *Pseudothermotoga* and *Petrotoga*) and induce viral production from one of these proviruses. This greatly expands the knowledge about viruses infecting the bacterial phylum Thermotogota, for which only three temperate viruses infecting *Marinitoga* species have been previously described. Moreover, we discovered that the identified Thermotogota proviruses may represent viruses of unusually broad host range that in some instances spans phyla. As a result, these viruses are likely major contributors to the extensive lateral gene transfer observed in Thermotogota and other members of deep biosphere.

Summary

Phylogenomic analyses of bacteria from the phylum Thermotogota have shown extensive lateral gene transfer with distantly related organisms, particularly with Firmicutes. One likely mechanism of such DNA transfer is viruses. However, to date only three temperate viruses have been characterized in this phylum, all infecting bacteria from the *Marinitoga* genus. Here we report 17 proviruses integrated into genomes of bacteria belonging to eight Thermotogota genera and induce viral particle production from one of the proviruses. All except an incomplete provirus from *Mesotoga*, fall into two groups based on sequence similarity, gene synteny and taxonomic classification. Proviruses of Group 1 are found in the genera *Geotoga*, *Kosmotoga*, *Marinitoga*, *Thermosipho* and *Mesoaciditoga* and are similar to the previously characterized *Marinitoga* viruses, while proviruses from Group 2 are distantly related to the Group 1 proviruses, have different genome organization and are found in *Petrotoga* and *Defluviitoga*. Genes carried by both groups are closely related to Firmicutes and Firmicutes (pro)viruses in phylogenetic analyses. Moreover, one of the groups show evidence of recent gene exchange and may be capable of infecting cells from both phyla. We hypothesize that viruses are responsible for a large portion of the observed gene flow between Firmicutes and Thermotogota.

Introduction

The phylum Thermotogota comprises anaerobic fermentative bacteria, most of which are thermophiles (Pollo *et al.*, 2015). They are common in subsurface environments such as marine vents, terrestrial hot springs and deep subsurface oil reservoirs (Bhandari and Gupta, 2014; Nesbø *et al.*, 2015, 2019; Foght *et al.*, 2017). On phylogenetic trees of 16S rRNA gene, Thermotogota is usually a deep branching bacterial lineage, while ribosomal proteins and other markers do not always agree with that placement (Zhaxybayeva *et al.*, 2009; Hug *et al.*, 2016). Such discrepancies are likely due to lateral gene transfer (LGT), which has been an important force shaping the genomes of Thermotogota, with Firmicutes and Archaea being their most notable gene transfer partners (Nelson *et al.*, 1999; Zhaxybayeva *et al.*, 2009; Pollo *et al.*, 2015). The LGT between Firmicutes and Thermotogota is so extensive that the two phyla have been suggested to be linked by ''highways of gene sharing'' (Zhaxybayeva *et al.*, 2009). However, how these inter-phylum gene-sharing events occur is still unclear.

The subsurface constitutes the largest biosphere on Earth and is estimated to contain ~70% of all cells (Magnabosco *et al.*, 2019). Viruses are commonly reported to be the most numerous biological entities on Earth (Paez-Espino *et al.*, 2016). Since some estimates suggest that up to 97% of all viruses on earth are found in soil and sediments (Anderson *et al.*, 2013; Cobián Güemes *et al.*, 2016), viruses are likely to be particularly important in subsurface environments. Moreover, although both prokaryotic cell and virus numbers decrease with depth, the virus-to-cell ratio can increase with depth (Anderson *et al.*, 2013; Engelhardt *et al.*, 2014; Walsh *et al.*, 2016). Phylogeographic studies of hyperthermophilic *Thermotoga* and mesophilic *Mesotoga* have revealed

genetic interaction between geographically distant populations, particularly among the hyperthermophilic *Thermotoga* (Nesbø *et al.*, 2015, 2019). Viruses are one potential source of such long-distance dispersal of genetic material (Ochman *et al.*, 2000), especially for anaerobic organisms where surface dispersal is problematic.

Although viruses are likely candidates for transferring DNA both within and between species, only three temperate siphoviruses (MCV1, MCV2, and MPV1), all infecting members of one Thermotogota genus, *Marinitoga*, have been described (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018). MCV1 and MCV2 infect *Marinitoga camini* strains isolated from deep-sea hydrothermal vents (Mercier *et al.*, 2018). MPV1 infects the deep-sea marine vent bacterium *Marinitoga piezophila*, where it is highjacked by a plasmid co-occurring in the same host, illustrating the potential route of gene mobilization in these ecosystems (Lossouarn *et al.*, 2015). The three viruses are found as proviruses in their host genomes and show similar genomic organization and virion morphology. Phylogenetic and protein sequence-similarity analyses of the viral ORFs revealed that they often group either with Firmicutes or Firmicutes' viruses, which suggests that viruses infecting members of Firmicutes and Thermotogota phyla share a common gene pool (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018).

Here we report 17 additional proviruses in Thermotogota genomes from eight Thermotogota genera, and a successful induction of one of these proviruses. The identified proviruses fall into two distinct groups. Both groups are closely related to Firmicutes viruses, and the proviruses from one of these groups may be able to infect cells from both phyla. We hypothesize that membrane transport proteins, such as ABC transporters, serve as receptors for Thermotogota viruses. We propose a mechanism that

could account for the highways of gene sharing observed between Thermotogota and Firmicutes, where LGT of viral genes encoding transmembrane proteins may make the host vulnerable to new viruses.

Material and Methods

Prediction and taxonomic classification of proviruses and functional annotation of their ORFs

One hundred eleven Thermotogota genomes were downloaded from either Genbank or IMG (Markowitz *et al.*, 2014) prior to June 2018. For draft genomes, the contigs were combined into 'artificially closed' genome using the "union" command from the EMBOSS package (version 6.6.0) (Rice *et al.*, 2000). Each genome was screened for the presence of proviruses using the Prophinder web server (Lima-Mendez *et al.*, 2008), PHAST web server (Zhou *et al.*, 2011), and PhiSpy (version 2.3) (Akhter *et al.*, 2012) between October 2014 and July 2018 (Supporting Table S1). These tools rely on genomic information of known bacterial viruses and will therefore be most successful in predicting dsDNA viruses since these are most numerous in the databases. For artificially closed genomes, the proviral regions that crossed contig borders were discarded. Putative provirus regions were inspected to identify the most likely provirus sequence by (1) looking for annotations indicating virus function (e.g. recombinases, capsid, tail) and/or annotations similar to those at the ends of the characterized viruses MPV1, MCV1 and MCV2 (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018); (2) identifying possible flanking tRNA genes; and (3) comparing the region to genomes from the same genus and defining the boundaries, to ensure that flanking genomic regions present in closely related

genomes without provirus were not included. We also looked for other features of proviruses such as strand switches, a relatively large gap between genes. See Table S1 for detailed notes on how the boundaries were selected for each provirus. Proviruses were considered complete if they contained modules for lysogeny, replication, packaging, head/tail morphogenesis and lysis. If one of these modules were missing, the provirus was scored as incomplete.

To further improve boundary information, we searched GenBank for close relatives of the proviruses ORFs using BLASTX (ver.2.2.26) (Camacho *et al.*, 2009) searches of the NCBI non-redundant (nr) database (O'Leary *et al.*, 2016) (accessed between July 2018 and March 2020). When homologs for > 4 ORFs from the predicted provirus were found in the same distantly related subject genome, the identified genome was downloaded and aligned to the Thermotogota genome carrying the provirus using Progressive Mauve (Darling *et al.*, 2010). The aligned region was used in combination with the virus prediction software to determine the boundaries of the provirus, as noted in Table S1.

Proviral ORF annotations were obtained from their respective GenBank entries and supplemented by results from BLASTP searches (Camacho *et al.*, 2009) of the *nr* database with an expected-value cutoff of 10^{-1}, and from HHpred searches (Zimmermann *et al.*, 2018) of the PDB database (Jones *et al.*, 2014) with a probability cutoff of 99%. In addition, recombinase- and terminase-encoding ORFs were annotated using InterProScan (Jones *et al.*, 2014), as implemented in Geneious v.10 (Biomatters Ltd.).

157 The sequences of the predicted proviruses were compared to each other using
158 BLASTN and TBLASTX (ver.2.2.26) (Camacho *et al.*, 2009) and visualized using
159 genoPlotR (Guy *et al.*, 2010) and Circos (Krzywinski *et al.*, 2009). Taxonomic
160 classification of the provirus genomes was carried out using searches of NCBI's viral
161 RefSeq database (v. 94) as implemented in VContact2 using Diamond (Buchfink *et al.*,
162 2015) to identify viral protein clusters and ClusterONE (Nepusz *et al.*, 2012) to obtain
163 virus clusters (Bolduc *et al.*, 2017; Bin Jang *et al.*, 2019). The resulting virus network
164 was drawn in Cytoscape (Shannon, 2003) including only viruses at most three nodes from
165 MPV1 and P8T1HF07V1. The VContact2 genome_by_genome summary file is available
166 on figshare (https://doi.org/10.6084/m9.figshare.15040044.v1). Taxonomic classification
167 was also assessed with Virfam (Lopes *et al.*, 2014), VIRIDIC (Moraru *et al.*, 2020) and
168 VIPTree (Nishimura *et al.*, 2017). Morphological classification was obtained with Virfam
169 (Lopes *et al.*, 2014).
170
171 **Inference of potential host range of the putative Thermotogota viruses.** A database
172 containing all proteins from 59 Thermotogota genomes without identified proviruses was
173 constructed in Geneious v.10. Translated Thermotogota provirus proteins (N=1,048) were
174 used as queries in BLASTP searches of this database. The provirus genes were scored as
175 present in the Thermotogota genomes if the query protein had a match with > 50% amino
176 acid identity and > 60% coverage.
177 CRPISPR spacer sequences from 90 Thermotogota genomes were obtained from
178 IMG (Markowitz *et al.*, 2014). The spacers were mapped to the provirus genomes in
179 Geneious v.10, allowing up to 10% nucleotide mismatches.

Phylogenetic analyses of provirus genes and candidate receptor genes.

Homologs of provirus genes selected for phylogenetic analysis were obtained by searching each translated proviral gene against *nr* database (accessed between December 2019 and March 2020), as well as a local database of all identified *Thermotogota* virus proteins, using BLASTP (version 2.2.26), with E-value cutoff of 10^{-1}. The 20 top-scoring matches from each database were retrieved and aligned using MAFFT v. 7.450 with the G-INS-I option (Katoh and Standley, 2013). Identical sequences and highly similar sequences from the same genus were removed. Alignment positions with > 50% gaps were trimmed. Phylogenetic trees were reconstructed using RAxML (Stamatakis, 2014) with WAG+G substitution model with four rate categories and 100 bootstrap replicates or using FastTree with JTT + G substitution model with 20 rate categories (Price *et al.*, 2010), as implemented in Geneious v.10.

Candidate receptor proteins in genomes of *Petrotoga* sp. 8T1HF07.NaAc.6.1, *Petrotoga olearia*, *Petrotoga mobilis*, *Petrotoga* sp. 9T1HF07.CasAA.8.2, *Defluviitoga tunisiensis*, *Lacticigenium naphtae*, and *Mahella australiensis*, which had proviruses assigned to Group 2 (see the Results section for definition), were identified in IMG using an amino acid identity cut-off of 50%. Homologs in *Geosporobacter ferrireducens* genome, which was not available in IMG, were identified using BLASTP search with E-value cutoff of 10^{-10}. Collection of additional homologs and phylogenetic analyses were carried out as described above.

Phylogenetic analysis of single copy gene in Thermotogota genomes available in GenBank (accessed May 27 2020) were done using the GToTree pipeline (Lee, 2019)

with the Bacterial HMM-set of 74 target genes. The resulting alignment was imported into Geneious Prime 2020.1.2 where sites with more than 50% gaps were removed, giving an alignment of 11,003 amino acid positions. The phylogenetic tree was reconstructed using FastTree with the JTT+G substitution model with 20 rate categories (Price *et al.*, 2010).

Virus induction and electron microscopy. *T. africanus* H17ap6033 and two *Petrotoga* isolates, *P.olearia* and *Petrotoga* sp. 8T1HF07.NaAc.6.1, were cultivated in a modified Ravot medium as previously described (Lossouarn *et al.*, 2015) at 65°C and 55°C, respectively. Attempts were made to increase the viral production of the strains by using mitomycin C, as reported previously (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018). 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 (Geslin *et al.*, 2003). 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. Five virions were used to measure tail and capsid size.

Results

Newly identified Thermotogota proviruses come from two distinct viral lineages. Analysis of 111 Thermotogota genomes identified 20 proviruses, including the three already characterized viruses from *Marinitoga* (MPV1, MCV1 and MCV2; Lossouarn *et al.*, 2015; Mercier *et al.*, 2018) and four likely remnants of proviruses (Supporting Table S1A and Supporting Table S2). One of the 20 proviruses is present with 100% nucleotide sequence identity in all six available *Thermosipho melanesiensis* genomes (Haverkamp *et al.*, 2018), and therefore is counted as just one novel provirus. An additional provirus was reported in a *Marinitoga lauensis* genome after we completed the screening (L'Haridon *et al.*, 2019). Due to its similarity to proviruses identified in other *Marinitoga* genomes, it was not included in our further analyses.

Based on genomic and sequence analyses the predicted proviruses can be divided into two distinct groups hereafter denoted as Group 1 (15 proviruses: 13 complete and 2 incomplete) and Group 2 (5 proviruses: 3 complete and 2 incomplete). First, the genome organization differs between the proviruses in two groups (Fig. 1 and Fig. 2). For instance, the lysogeny module is found at the 5' end of the Group 1 proviruses, and near the 3' end of the Group 2 viruses. Moreover, the lysogeny module is encoded on the opposite strand compared to most other genes in the Group 1 proviruses, this is not the case for the Group 2 proviruses. Second, the genes within each group are more similar

than the genes between groups (Fig. 3, Supporting Table S2). Third, the two groups form separate clusters in the VContact2 network (Supporting Fig. S1, panel A). Finally, the two groups show up as different clades on the Viral Proteomic Tree (Supporting Fig. S2). Group 1 proviruses are found in the genera *Marinitoga*, *Thermosipho*, *Kosmotoga*, *Geotoga* and *Mesoaciditoga*, while Group 2 proviruses are limited to the genera *Petrotoga* and *Defluviitoga* (Supporting Fig. S1, panel B). However, the presence of 29 protein families (representing 10% and 24% of the proteins from Group 1 and 2, respectively) shared between the two groups (Supporting Table S3) suggests that LGT may occur between the viruses of the two groups, or that the two groups share a distant common ancestor.

Both the Group 1 and Group 2 proviruses are likely to have a siphovirus morphology based on their head, neck and tail gene sequences (Lopes *et al.*, 2014), and the morphology observed for the earlier characterized MPV1, MCV1, MCV2 viruses (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018) and the TAV1 virus induced in the current study (see below). The incomplete provirus from *Mesotoga* sp. SC_TOLDC showed highest sequence similarity to the Group 1 viruses (Fig. 3, Supporting Table S2), but was predicted to have myovirus morphology (Lopes *et al.*, 2014). This provirus sequence was therefore included in Group 1 for sequence comparisons only.

Figure 1

Figure 2

Figure 3

Classification and genomic features of Group 1 proviruses.

None of the proviruses have significant nucleotide identity with viral genomes in the NCBI nr/nt database. Several recent studies have focused on the best ways to classify bacterial virus species, genera and higher taxonomic ranks (Gregory *et al.*, 2016; Bin Jang *et al.*, 2019; Barylski *et al.*, 2020). Following the latest proposed taxonomic criteria (Turner *et al.*, 2021), where viruses with > 70% nucleotide identity over the full genome belong to the same genus and viruses with > 95% nucleotide identity belong to the same species, the VIRIDIC tool (Moraru *et al.*, 2020) indicated that the complete proviruses could be assigned to 13 new candidate species and 13 new candidate genera (Supporting Table S4). The network analysis and sequence similarity searches suggest that the closest relatives of the Group 1 proviruses are Firmicutes' viruses, since 15% of the provirus genes have Firmicutes as the top-scoring match, if members of the proviruses' host genus are excluded (Supporting Fig. S1, panel A and Table S5).

The proviruses have the same modular structure as genomes of the earlier described MPV1, MCV1 and MCV2 viruses (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018) (Fig. 1). The 5' module contains genes involved in lysogeny and is encoded on the opposite strand compared to the rest of the virus genes. The lysogeny module is followed by modules for replication, packaging, morphogenesis and host lysis. Similar to the described *Marinitoga* viruses (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018), the gene content of lysogeny module of all examined proviruses is very variable, with only the recombinase gene conserved (Fig. 1).

All Group 1 proviruses are inserted next to a tRNA gene. Eight of them (the *Marinitoga* proviruses MPV1, MCV1, MCV2, MHV1 and M1137V2; the *Kosmotoga pacifica* provirus KPV1, *Geotoga petrae* provirus GPV1, and *Mesoaciditoga lauensis* provirus MLV1), are inserted next to the tRNA-Glu gene, and carry similar site-specific DNA serine recombinases (KEGG Orthology; K06400, homologs of Marpi_0291 in MPV1 from *M. piezophila*) (Fig. 4). The *Thermosipho* proviruses TMV1 and T1074V1 carry more distant homologs of this recombinase (Fig. 4) and are inserted next to the tRNA-Phe gene. The most divergent homolog of the serine recombinase is present in TAV1, which is inserted next to the tRNA-Pro gene. We hypothesize that these recombinases are integrases that specifically recognize the tRNA genes next to which the provirus is inserted (i.e., the tRNA-Glu, tRNA-Phe, or tRNA-Pro genes). The remaining three proviruses (M1135V1, M1138V1, and M1137V1) are inserted next to a tRNA-Cys gene and may also use similar integration mechanism, but these proviruses do not have a detectable serine recombinase homologs. M1135V1 and M1138V1 have homologous ORFs of unknown function in the recombinase gene position (Fig. 1), which may or may not provide this function.

Figure 4

Another typical viral protein that shows variation across the Group 1 proviruses is the large subunit of the terminase protein involved in the packaging of viral DNA into the virus particle (Catalano, 2000). The terminases from the Group 1 proviruses show high levels of diversity and fall into three clusters of proteins with > 25 % identity and

317 BLASTP E-value < 0.01 (Supporting Fig. S3). The first cluster, exemplified by the
318 protein in MPV1 (Marpi_0320), contains a PBSX family domain and has close homologs
319 in MCV1 and MCV2 and the proviruses in *Marinitoga* sp. 1135, *Marinitoga* sp. 1138,
320 *Thermosipho* sp. 1074 and *M. lauensis*. The second, exemplified by the terminase in
321 TAV1 (H17ap60334_04902), contains a 'Terminase_lsu_T4-like' domain and has close
322 homologs in the proviruses from *K. pacifica*, *T. melanesiensis*, *Mesotoga* sp. TolDC and
323 *G. petrae*. The third cluster is found in the proviruses in *M. hydrogenitolerans* and
324 *Marinitoga* sp. 1137 (BUA62_RS02495, LN42_01905 and LN42_00550). These
325 terminases also contain a 'Terminase_lsu_T4-like' domain.
326 In addition to the recombinase and terminase, other typical viral proteins such as
327 tail tape measure, capsid and portal proteins were identified, but did not always show
328 detectable similarity among the proviruses (Fig. 1). Two transcription regulators
329 (Marpi_0297 and Marpi_0298 in MPV1), a DNA repair exonuclease (Marpi_0340 in
330 MPV1) and a single stranded DNA-binding protein (Marpi_0306 in MPV1), show
331 relatively high identity (32 - 100 %) across most Group 1 proviruses (Fig. 1, Supporting
332 Table S4). Genes encoding two hypothetical proteins (homologs of Marpi_0299 and
333 Marpi_0338 in MPV1) are shared among 10 of the proviruses (36-96% identity),
334 suggesting these genes may provide important viral functions.
335
336 **Broad host range of Group 1 proviruses.**
337 Detection of Group 1 proviruses in the genera *Marinitoga*, *Thermosipho*, *Kosmotoga*,
338 *Geotoga* and *Mesoaciditoga*, suggests that the Group 1 viruses are widespread among
339 Thermotogota, particularly among organisms inhabiting hydrothermal vents (Supporting

340 Table S1). Such wide distribution and relatively high sequence identity among the
341 proviral genomes suggest that the Group 1 temperate viruses might have broad host
342 ranges (Fig. 1, Supporting Table S4). Experiments showing that MPV1 from *M.*
343 *piezophila* can infect and transfer a plasmid to a *Thermosipho* isolate is consistent with
344 this hypothesis (Lossouarn *et al.*, 2015).
345 Further support comes from mapping of CRISPR spacer sequences from 90
346 Thermotogota genomes to the Group 1 proviruses. Five of the 17 proviruses matched
347 CRISPR spacers in the genomes from a different genus (Table 1). For example, the
348 *Thermosipho* provirus TAV1 had 35 matches to spacers in the genomes of
349 *Pseudothermotoga* and *Thermotoga* spp. Anecdotal evidence corroborates an ability of
350 TAV1 to infect *Thermotoga* spp. Back in 2005, when the sample from the Hibernia oil
351 reservoir containing TAV1 and its host, *T. africanus* H17ap60334, was being processed
352 by one of us (Camilla L. Nesbø) in the laboratory, *Thermotoga* isolates from Troll oil
353 reservoir in the North Sea, which were at the same time being transferred to fresh media,
354 experienced a mass death. Analysis of the genomes of the surviving *Thermotoga* isolates
355 (Nesbø *et al.*, 2015) revealed presence of three CRISPR spacer matching the TAV1
356 genome (Supporting Fig. S4). These spacers were located in the middle of the CRISPR
357 arrays, indicating that they were likely not new acquisitions (Rath *et al.*, 2015). Therefore
358 the only surviving isolates of *Thermotoga* likely had already experienced and survived
359 TAV1 or related virus infections in the oil reservoir.
360
361 Table 1
362

Classification, genomic features and distribution of Group 2 proviruses.

Group 2 consists of three complete proviruses in the genomes of *Petrotoga* sp. 8T1HF07 (P8T1HF07V1), *Petrotoga olearia* (POV1) and *Defluviitoga tunesiensis* (DTV1) (Fig. 2), and two incomplete proviruses in the genomes of *Petrotoga mobilis* SJ95 and *Petrotoga* sp. 9T1HF07 (Supporting Fig. S5). Due to the short length of the incomplete proviruses, they were not included in the remaining analyses of this section.

Following the taxonomic classifications criteria described above, the three complete proviruses P8T1HF07V1, DTV1 and POV1 are assigned to three new viral species (Supporting Table S4). Based on the head-neck-tail module classification (Lopes *et al.*, 2014), these proviruses likely display a siphovirus morphology of Type1 - Cluster 2. All hosts of the previously described members of this *Siphoviridae* lineage belong to Firmicutes (Lopes *et al.*, 2014). In agreement with this, similarity searches revealed that these proviruses show very high similarity to proviruses of three Firmicutes genomes: *Lacticigenium naphta*e (LNV1), *Geosporobacter ferrireducens* (GFV1) and *Mahella australiensis* (MAV1) (Fig. 2, Supporting Table S1B, Supporting Table S4).

The sequences and genome organization of the three complete Group 2 proviruses differ considerably from that of Group 1 (Fig. 1 and Fig. 2). These proviruses are also not located next to tRNA genes. The 5' module encodes genes involved in virus replication and transcription, and the comparative genomic analysis shows high level of diversity in this region (Fig. 2). This module is followed by highly conserved packaging, morphogenesis and lysis modules. The lysogeny module is located at the 3' end of the provirus. The site-specific serine recombinases carried by the Group 2 proviruses in this module are distant homologs of the earlier discussed Group 1 recombinases (Fig. 4).

When comparing the Group 2 provirus genomes from Thermotogota and Firmicutes, each Thermotogota provirus is more similar to a Firmicutes provirus than to other Thermotogota proviruses (Fig. 2, Supporting Table S4)**.** Alignments of the two Thermotogota-Firmicutes provirus pairs, P8T1HF07V1 and GFV1, and DTV1 and MAV1, have 66.7 % and 57.7 % intergenomic similarity values, respectively (Supporting Table S4). Moreover, P8T1HF07V1 has 97% nucleotide identity to GFV1 over the specific subregion that encode structural genes, such as the tail fibers used by the virus to adsorb to the host cell, and the genes for DNA packaging and genome integration (Fig. 2). Similarly, the same regions in DTV1 and MAV1 have 95-97% identity. In contrast, the same regions in P8T1HF07V1 and POV1, and P8T1HF07V1 and DTV1 have 53 and 75% nucleotide identity, respectively. Such similarity patterns suggest that these viruses may be able to infect hosts from both Thermotogota and Firmicutes.

In contrast to the Group 1, none of the Group 2 proviruses had matches to CRISPR spacers in 90 Thermotogota genomes, suggesting that the Group 2 viruses have a more restricted host range within the Thermotogota or started to infect members of this phylum recently.

Successful induction of TAV1 from *T. africanus* H17ap60333.

Induction assays were performed on three of the putatively lysogenized *Thermotogota*: *T. africanus* H17ap6033 (Group 1), *Petrotoga* sp. P8T1HF07 (Group 2) and *P. olearia* (Group 2). Only the provirus in *T. africanus* H17ap6033 (TAV1) was successfully induced using mitomycin C. 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 sequence-based classification. TAV1 morphology is similar to the three previously characterized temperate *Marinitoga* viruses, whose virion tails were just slightly longer (Lossouarn *et al.*, 2015; Mercier *et al.*, 2018). In addition to viral particles, a release of membrane vesicles or toga fragments was regularly observed (Fig. 5b).

Figure 5

While the induction of the proviruses in *Petrotoga* sp. 8T1HF07 (P8T1HF07V1) and *P. olearia* (POV1) using mitomycin C was unsuccessful, membrane vesicles of various sizes and shapes (20 - 100nm) were produced by the cells, and in particular by the induced *Petrotoga* sp. 8T1HF07 cells. Analysis of the supernatant of the latter culture revealed similarly-sized round-shaped vesicles connected together in long chains by hooking onto 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 a remaining cell (Fig. 5d). It is unknown if the provirus or stressors influence the production of these vesicles and structures, or if they are produced spontaneously.

A potential receptor for the Group 2 viruses

Several types of structures on the surface of bacteria, such as membrane proteins, flagella, pili, or carbohydrate moieties, can act as virus receptors (Stone *et al.*, 2019).

Most siphoviruses of Gram-negative bacteria, and some of Gram-positive bacteria, use proteinaceous receptors for adsorption (Bertozzi Silva *et al.*, 2016; Zhang *et al.*, 2020). If the Group 2 viruses use the same protein receptor to attach to both Thermotogota and Firmicutes cells, the large phylogenetic distance between these hosts offers an opportunity to identify possible membrane protein receptors bioinformatically, since the receptor proteins would be expected to be conserved across the genomes from both phyla. It should be noted that this approach would only identify possible protein receptors, while potential shared carbohydrate receptors would not be detected.

Four predicted membrane proteins with transmembrane helices were identified in all genomes carrying a Group 2 provirus. One of these was the viral holin gene, leaving three receptor 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 and the oxaloacetate decarboxylase homologs are widely distributed among Thermotogota (Supporting Fig. S6). In contrast, the ABC transporter is, among the Thermotogota, restricted to *Petrotoga* and *Defluviitoga*, the two genera where the Group 2 proviruses are observed (Supporting Fig. S6, panel C). Moreover, the phylogenetic analysis suggests the homologs in *Petrotoga* and *Defluviitoga* originated from an LGT event with a Firmicute (Supporting Fig. S6**)**. These proteins show particularly high amino acid sequence similarity in the C-terminal domain of both Thermotogota and Firmicutes homologs, which is facing the exterior of the cell and could serve as a virus target (Supporting Fig. S7). Although experiments are needed to demonstrate if any of these proteins functions as receptor for these viruses, we suggest

454 that the ABC-transporter ATP-binding protein is a strong candidate for a Group 2 virus
455 receptor.
456
457 **Moron genes are abundant in the identified proviruses.**
458 Many temperate viruses are known to carry moron genes, which are genes that are not
459 conserved across virus genomes and do not have a typical viral function (Juhala *et al.*,
460 2000; Cumby *et al.*, 2012; Taylor *et al.*, 2019). Some morons may increase the fitness the
461 host cell and/or the virus. The detected proviruses of Thermotogota are no exception:
462 based on annotation the Group 1 proviruses carry up to 6 likely morons (Fig. 1), while the
463 Group 2 proviruses have between 4 and 13 potential morons (Fig. 2). However, it should
464 be noted that because the 5' ends were hard to define for Group 2 proviruses, some of the
465 moron genes at the 5' ends might not be part of the proviruses. Sequencing virus DNA
466 isolated from capsids will help resolve this issue in the future.
467 Among the morons are several proteins that may confer a selective advantage to
468 the host (Fig. 1 and Fig. 2, Supporting Table S2). For instance, M1138V1 carry two
469 genes involved in sulfur metabolism. The Group 2 proviruses encode several transporters,
470 peptidases and hydrolases, likely to be beneficial for these heterotrophic bacteria. In
471 addition, all proviruses carry several hypothetical proteins that may also have non-viral
472 functions.
473
474 **Evidence for the *Thermotogota* viruses' impact on lateral gene transfer**
475 Eight hundred seventy homologs of 106 proviral genes (10% of all the proviral genes)
476 were detected in 54 out of 59 Thermotogota genomes with no detectable proviruses

(Supporting Table S6). It should be noted that some provirus genes, e.g. the Group 1 recombinases and terminases (Fig. 4, Supporting Fig. S3), did not pass our stringent screening criteria (see Material and Methods), thus these represent minimum estimates of matches to proviral genes in these genomes. Notably, 370 of 870 were homologs of 28 moron genes, suggesting that the viruses may facilitate exchange of "host" genes among Thermotogota. Moron genes also had the highest number of homologs across the 54 genomes, with the most abundant being a queuine tRNA-ribosyltransferase in the Group 2 provirus DTV1 (found in 48 genomes) and an aldo/keto reductase in Group 1 provirus GPV1 (found in 41 genomes) (Supporting Table S6).

Among phylogenetically informative datasets, 10 proviral genes group within Thermotogota and 15 were likely acquired from Firmicutes, suggesting that many of the proviral genes originated either in Thermotogota and Firmicutes (Supporting Table S6 and Fig. S9**)**. For instance, the above-described abundant moron gene queuine tRNA-ribosyltransferase is of Thermotogota origin (Supporting Fig. S9f), while the aldo/keto reductase appears to be of Firmicutes origin (Supporting Fig. S9b**)**. In the phylogeny of another moron gene, a cadmium or heavy metal transporter found in Firmicutes provirus MAV1 and Thermotogota Group 2 proviruses DTV1 and POV1, the provirus genes group closely with Firmicutes' homologs (Supporting Fig. S8). Notably, the other closely related Thermotogota homologs are found in three *Fervidobacterium* and *Pseudothermotoga* species, genera where no proviruses have yet been identified. Inspecting the genomic region surrounding these genes in *Fervidobacterium* and *Pseudothermotoga*, revealed that the homolog of the proviral recombinase (Fig. 4) is located immediately upstream of the transporter gene. No other typical virus genes were

observed in these regions. However, their genomic proximity together with the phylogenetic analyses (Fig. 4, Supporting Fig. S8), suggests these genes are likely remnants of proviruses. This also indicates that viruses related to Group 2 proviruses may have broader host range than we presently detect.

Taken together the above analyses suggest that the viruses of both Group 1 and Group 2 may facilitate exchange of genes not only among Thermotogota, but also between Thermotogota and Firmicutes.

Discussion

In our search for proviruses in genomes of Thermotogota, we discovered two distinct groups of temperate siphoviruses that have lysogenized this bacterial phylum. These proviruses may represent multiple new viral species and genera. Our analyses suggest that these viruses likely have broad host range that spans at least multiple genera. We also found that the identified proviruses lineages are closely related to Firmicutes' viruses. In agreement with this, recent studies indicate that although narrow host ranges appear to be most common, broad-host-range viruses are more common in nature than previously thought with some viruses infecting cells assigned to different higher taxonomic ranks, even phyla (Paez-Espino *et al.*, 2016; de Jonge *et al.*, 2019).

One of the Group 1 proviruses (TAV1) was induced and shown to produce virus particles. The provirus resides in a genome of a *T. africanus* isolate from the Hibernia oil reservoir off the Canadian east coast. The analysis of CRISPR spacers suggested that this virus may have a particularly wide host range, with the highest number of spacer-matches in genomes from outside its host's genus (Table 1). For instance, a virus very similar to

TAV1 had likely infected *Thermotoga* spp. isolates from the North Sea Troll oil reservoir (Supporting Fig. S7). Similar predatory virus pressure in geographically and geologically remote subsurface environments have been observed for *Methanohalophilus* isolates from reservoirs in the USA and Russia (Borton *et al.*, 2018).

We were not able to induce virus production from the selected Group 2 proviruses This could be due to these proviruses currently being inactive, or we may not have applied the right conditions to induce the expression of these proviruses. Nevertheless, the high level of sequence identity between Group 2 provirus sequences from Thermotogota and Firmicutes phyla suggests that they have been active very recently.

Many of the genes carried by both Group 1 and Group 2 proviruses are found in genomes of Thermotogota that do not have detectable proviruses. These genes often display closest sequence similarity to Firmicutes or viruses that infect Firmicutes, and many can be classified as morons. This suggests that both Group 1 and Group 2 viruses transfer genes within Thermotogota and between Thermotogota and Firmicutes, and may serve as a major mechanism for the earlier reported large amounts of lateral gene transfer between Thermotogota and Firmicutes (Nelson *et al.*, 1999; Zhaxybayeva *et al.*, 2009).

Based on our bioinformatic and phylogenetic analyses, we propose that an ABC transporter may serve as a receptor for at least some of these proviruses. ABC transporter proteins are, to our knowledge, not commonly identified as viral receptors. However, *Lactoccocus* viruses from the siphoviral c2 group have been shown to use membrane proteins Pip or YjaE, both with sequence similarity to ABC-transporter domains, as secondary receptors (Millen and Romero, 2016; Stone *et al.*, 2019).

Intriguingly, transporters, and ABC transporters in particular, are among the most frequently transferred genes both within the Thermotogota and between Thermotogota and Firmicutes (Nelson *et al.*, 1999; Nesbø *et al.*, 2002; Zhaxybayeva *et al.*, 2009). Transporter genes were also detected in the provirus genomes. The possibility that transporters can function as viral receptors in the Thermotogota therefore suggests that acquiring a new transporter, perhaps via a viral 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 but could also make viruses the vehicles on the highways of gene sharing observed between the Thermotogota and Firmicutes.

Genes encoding proteins for membrane transport, including ABC transporters, have been observed in large eukaryotic viruses and *Caudovirales* viruses (Greiner *et al.*, 2018), and thus the proposed process could operate widely among bacteria. This is contrary to a role commonly assigned to morons where they often confer resistance to infections by other viruses (Taylor *et al.*, 2019). 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 viruses potentially infecting organisms from different phyla further demonstrates that viruses are key actors in the evolution of microbial diversity.

Acknowledgements

This work is supported by a Research Council of Norway award (project no. 180444/V40) to C.L.N., by the Sino-French LIA/PRC 1211 MicrobSea to J.L. and by the Simons Foundation Investigator in Mathematical Modeling of Living Systems award 327936 to O.Z. Strains were obtained from the Université de Bretagne Occidentale Culture Collection (UBOCC, Plouzané, France, www.univ-brest.fr/ubocc).

Conflict of Interest Statement:

The authors declare no conflict of interest.

References

Akhter, S., Aziz, R.K., and Edwards, R.A. (2012) PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Res* **40**: e126–e126.

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2013) The Deep Viriosphere: Assessing the Viral Impact on Microbial Community Dynamics in the Deep Subsurface. *Rev Mineral Geochem* **75**: 649–675.

Barylski, J., Enault, F., Dutilh, B.E., Schuller, M.B., Edwards, R.A., Gillis, A., et al. (2020) Analysis of Spounaviruses as a Case Study for the Overdue Reclassification of Tailed Phages. *Syst Biol* **69**: 110–123.

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

Bhandari, V. and Gupta, R.S. (2014) The Phylum Thermotogae. In *The Prokaryotes*. Rosenberg, E., DeLong, E.F., Lory, S., Stackebrandt, E., and Thompson, F. (eds). Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 989–1015.

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

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

Borton, M.A., Daly, R.A., O'Banion, B., Hoyt, D.W., Marcus, D.N., Welch, S., et al. (2018) Comparative genomics and physiology of the genus *Methanohalophilus* , a prevalent methanogen in hydraulically fractured shale. *Environ Microbiol* **20**: 4596–4611.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Catalano, C.E. (2000) The terminase enzyme from bacteriophage lambda: a DNA-packaging machine. *Cell Mol Life Sci CMLS* **57**: 128–148.

Cobián Güemes, A.G., Youle, M., Cantú, V.A., Felts, B., Nulton, J., and Rohwer, F. (2016) Viruses as Winners in the Game of Life. *Annu Rev Virol* **3**: 197–214.

Cumby, N., Davidson, A.R., and Maxwell, K.L. (2012) The moron comes of age. *Bacteriophage* **2**: e23146.

Darling, A.E., Mau, B., and Perna, N.T. (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE* **5**: e11147.

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

Foght, J.M., Gieg, L.M., and Siddique, T. (2017) The microbiology of oil sands tailings: past, present, future. *FEMS Microbiol Ecol* **93**:.

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

Gregory, A.C., Solonenko, S.A., Ignacio-Espinoza, J.C., LaButti, K., Copeland, A., Sudek, S., et al. (2016) Genomic differentiation among wild cyanophages despite widespread horizontal gene transfer. *BMC Genomics* **17**: 930.

Greiner, T., Moroni, A., Van Etten, J., and Thiel, G. (2018) Genes for membrane transport proteins: Not so rare in Viruses. *Viruses* **10**: 456.

Guy, L., Roat Kultima, J., and Andersson, S.G.E. (2010) genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**: 2334–2335.

Haverkamp, T.H.A., Geslin, C., Lossouarn, J., Podosokorskaya, O.A., Kublanov, I., and Nesbø, C.L. (2018) Thermosipho spp. immune system differences affect variation in genome size and geographical distributions. *Genome Biol Evol*.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., et al. (2016) A new view of the tree of life. *Nat Microbiol* **1**: 16048.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014) InterProScan 5: genome-scale protein function classification. 5.

de Jonge, P.A., Nobrega, F.L., Brouns, S.J.J., and Dutilh, B.E. (2019) Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends Microbiol* **27**: 51–63.

Juhala, R.J., Ford, M.E., Duda, R.L., Youlton, A., Hatfull, G.F., and Hendrix, R.W. (2000) Genomic sequences of bacteriophages HK97 and HK022: pervasive genetic mosaicism in the lambdoid bacteriophages 1 1Edited by M. Gottesman. *J Mol Biol* **299**: 27–51.

Katoh, K. and Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**: 772–780.

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

Lee, M.D. (2019) GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* **35**: 4162–4164.

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

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

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

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

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

Magnabosco, C., Biddle, J.F., Cockell, C.S., Jungbluth, S.P., and Twing, K.I. (2019) Biogeography, Ecology, and Evolution of Deep Life. In *Deep Carbon*. Orcutt, B.N., Daniel, I., and Dasgupta, R. (eds). Cambridge University Press, pp. 524–555.

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

Mercier, C., Lossouarn, J., Nesbø, C.L., Haverkamp, T.H.A., Baudoux, A.C., Jebbar, M., et al. (2018) Two viruses, MCV1 and MCV2, which infect Marinitoga bacteria isolated from deep-sea hydrothermal vents: functional and genomic analysis. *Environ Microbiol* **20**: 577–587.

Millen, A.M. and Romero, D.A. (2016) Genetic determinants of lactococcal C2viruses for host infection and their role in phage evolution. *J Gen Virol* **97**: 1998–2007.

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

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., et al. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. *Nature* **399**: 323–329.

Nepusz, T., Yu, H., and Paccanaro, A. (2012) Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* **9**: 471–472.

Nesbø, C.L., Charchuk, R., Pollo, S.M.J., Budwill, K., Kublanov, I.V., Haverkamp, T.H.A., and Foght, J. (2019) 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* **21**: 456–470.

690 Nesbø, C.L., Nelson, K.E., and Doolittle, W.F. (2002) Suppressive Subtractive
691 Hybridization Detects Extensive Genomic Diversity in Thermotoga maritima. *J*
692 *Bacteriol* **184**: 4475–4488.
693 Nesbø, C.L., S Swithers, K., Dahle, H., Haverkamp, T.H., Birkeland, N.-K., Sokolova,
694 T., et al. (2015) Evidence for extensive gene flow and Thermotoga subpopulations
695 in subsurface and marine environments. *ISME J* **9**: 1532–1542.
696 Nishimura, Y., Yoshida, T., Kuronishi, M., Uehara, H., Ogata, H., and Goto, S. (2017)
697 ViPTree: the viral proteomic tree server. *Bioinformatics* **33**: 2379–2380.
698 Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000) Lateral gene transfer and the
699 nature of bacterial innovation. **405**: 6.
700 O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., et al.
701 (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic
702 expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–D745.
703 Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M.,
704 Mikhailova, N., et al. (2016) Uncovering Earth's virome. *Nature* **536**: 425–430.
705 Pollo, S.M.J., Zhaxybayeva, O., and Nesbø, C.L. (2015) Insights into thermoadaptation
706 and the evolution of mesophily from the bacterial phylum *Thermotogae*. *Can J*
707 *Microbiol* **61**: 655–670.
708 Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 – Approximately Maximum-
709 Likelihood Trees for Large Alignments. *PLoS ONE* **5**: e9490.
710 Rath, D., Amlinger, L., Rath, A., and Lundgren, M. (2015) The CRISPR-Cas immune
711 system: Biology, mechanisms and applications. *Biochimie* **117**: 119–128.
712 Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: The European Molecular
713 Biology Open Software Suite. *Trends Genet* **16,**: 276–277.
714 Shannon, P. (2003) Cytoscape: A Software Environment for Integrated Models of
715 Biomolecular Interaction Networks. *Genome Res* **13**: 2498–2504.
716 Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-
717 analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
718 Stone, E., Campbell, K., Grant, I., and McAuliffe, O. (2019) Understanding and
719 Exploiting Phage–Host Interactions. *Viruses* **11**: 567.
720 Taylor, V.L., Fitzpatrick, A.D., Islam, Z., and Maxwell, K.L. (2019) The Diverse Impacts
721 of Phage Morons on Bacterial Fitness and Virulence. *Adv Virus Res* **103**: 1–31.
722 Turner, D., Kropinski, A.M., and Adriaenssens, E.M. (2021) A Roadmap for Genome-
723 Based Phage Taxonomy. *Viruses* **13**: 506.
724 Walsh, E.A., Kirkpatrick, J.B., Pockalny, R., Sauvage, J., Spivack, A.J., Murray, R.W., et
725 al. (2016) Relationship of Bacterial Richness to Organic Degradation Rate and
726 Sediment Age in Subseafloor Sediment. *Appl Environ Microbiol* **82**: 4994–4999.
727 Zhang, Z., Yu, F., Zou, Y., Qiu, Y., Wu, A., Jiang, T., and Peng, Y. (2020) Phage protein
728 receptors have multiple interaction partners and high expressions. *Bioinformatics*
729 **36**: 2975–2979.
730 Zhaxybayeva, O., Swithers, K.S., Lapierre, P., Fournier, G.P., Bickhart, D.M., DeBoy,
731 R.T., et al. (2009) On the chimeric nature, thermophilic origin, and phylogenetic
732 placement of the Thermotogales. *Proc Natl Acad Sci* **106**: 5865–5870.
733 Zhou, Y., Liang, Y., Lynch, K.H., Dennis, J.J., and Wishart, D.S. (2011) PHAST: A Fast
734 Phage Search Tool. *Nucleic Acids Res* **39**: W347–W352.

735 Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., et al. (2018)
736 A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred
737 Server at its Core. *J Mol Biol* **430**: 2237–2243.
738
739

740

741 **Table 1. CRISPR spacer matches to provirus genomes in Thermotogota genomes.**

742 Matches to spacers from the provirus' host genome are labeled as a self-match.

provirus name	Genome with a CRISPR spacer match (number of spacers). Same genus as (pro)virus host.	Genome with a CRISPR spacer match (number of spacers). Different genus than (pro)virus host.
KPV1	*Kosmotoga olearia* (1)	
MHV1	*Marinitoga* sp. 1154 (1)	*Pseudothermotoga elfii* NBRC107921 (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	*Marinitoga* sp. 1135 (2)	*Pseudothermotoga elfii* NBRC107921 (1)
M1137V2	*Marinitoga piezophila* (1)	
M1138	*Marinitoga* sp. 1137 (1)	*Thermosipho africanus* TCF52B (1), *Thermosipho melanesiensis* (1), *Pseudothermotoga elfii* NBRC107921 (1)

TAV1*	*Thermosipho africanus* H17ap60334 (3) self -match, *Thermosipho africanus* TCF52B (2), *Thermosipho africanus* Ob7 (1), *Thermosipho 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), *Thermotoga* sp. MC24 (2), *Pseudothermotoga elfii lettingae* (4), *Pseudothermotoga elfii* NBRC107921 (9), *Pseudothermotoga elfii* DSM9442 (2)
TMV1	*Thermosipho melanesiensis* (1) self-match, *Thermosipho africanus* TCF52B (1), *Thermosipho africanus* Ob7 (1)	*Thermotoga* sp. TBGT1765 (1), *Thermotoga* sp. Mc24 (1), *Pseudothermotoga elfii* NBRC107921 (1), *Pseudothermotoga elfii lettingae* (2)
T1074V1	*Thermosipho affectus* Bl1070 (1), *Thermosipho affectus* 1223 (3), *Thermosipho affectus* Bl1063 (1)	

743 *Proviruses that have been induced and shown to produce virus particles.
744

745

Fig. 1. Comparison of sequences from all detected Group 1 proviruses. Provirus name and the species of its host are shown to the left of the nucleotide sequence, in which predicted ORFs are depicted as arrows. Proviruses that have been induced and shown to produce virus particles are marked with an asterisk and the three previously characterized *Marinitoga* viruses are shown in bold font. The lines connect regions of adjacent viruses that have TBLASTX similarity of more than 30% over 100bp. Lines are colored in red or blue indicating that the matching sequences are encoded in the same or opposite strand, respectively. The ORFs are color-coded based on their predicted function and should be considered approximate. Selected gene annotations are included and abbreviated as follows. Ser recomb: serine recombinase, LexA: LexA repressor, ParB: ParB-like nuclease, RecT: RecT family recombinase, nucl hydrolase: p-loop triphosphate nucleoside hydrolase, 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, Ad1: adaptor protein Ad1, Hc1: head closure protein Hc1, Ne1 : neck protein Ne1, Tc1: tail completion Tc1, mtp: major tail 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,

769 metal bp: metal-binding protein, CMP hydrolase: cytidine 5′-monophosphate hydrolase.
770 Head-to-tail indicates the head-to-tail connector proteins. The figure was produced using
771 genoPlotR (Guy *et al.*, 2010).
772
773 **Fig. 2. Comparison of sequences from three complete Thermotogota Group 2**
774 **proviruses and their Firmicutes' homologs.** Provirus name (in red for Thermotogota
775 and blue for Firmicutes) and the species of its host are shown to the left of the nucleotide
776 sequence, in which predicted ORFs are depicted as arrows. The lines connect regions of
777 adjacent viruses that have TBLASTX similarity of more than 30% over 100bp. Lines are
778 colored in red or blue indicate that the matching sequences encoded in the same or
779 opposite strand, respectively. The ORFs are color-coded based on their predicted function
780 and should be considered approximate. Selected gene annotations are included and
781 abbreviated; HMT ATPase: heavy metal translocating ATPase, FMN reductase: flavine
782 mono nucleotide reductase, HAD family phosphatase: haloacid dehalogenase superfamily
783 of hydrolase). Head-to-tail indicates the head-to-tail connector proteins. Connector genes
784 predictions are the same for all the proviruses, specifically from 5' to 3' an adaptor
785 protein Ad1, a head closure protein Hc1 and a neck protein Ne1 which are followed by a
786 tail completion Tc1. The figure was produced using genoPlotR (Guy *et al.*, 2010).
787
788 **Fig. 3. Comparison of representative Thermotogota proviruses**. Due to sequence
789 similarity, only one provirus per Thermotogota genus is shown. The nucleotide sequences
790 of the proviruses are arranged around the circle and color-coded. Numbers indicate
791 kilobases (kb) and grey boxes outline locations of predicted genes. Lines connecting

792 different proviral sequences represent TBLASTX matches between the proviral regions,
793 with the percent identity shown in histograms at the ends of each line on a scale from 0 to
794 100%. The plot was created using Circos (Krzywinski *et al.*, 2009).
795
796 **Fig. 4. Maximum likelihood tree of recombinases found in Thermotogota proviruses**
797 **and of their homologs in Firmicutes proviruses, and Thermotogota and Firmicutes**
798 **genomes.** Host names of Thermotogota and Firmicutes proviruses are colored in red and
799 blue, respectively. The names of their proviruses are added next to the host name. Names
800 of Thermotogota homologs that either resided outside of proviral regions or come from a
801 genome without detected proviruses are shown in black. Branches without labels
802 represent Firmicutes without an identified Group 2 provirus. Homologs from incomplete
803 proviruses are labeled with "(in)". Circles on the branches represent bootstrap support,
804 and only values above 70% are shown. Some proteins have identical amino acid
805 sequences in more than one organism. The protein labelled 'Bacteria inc. POLV1
806 *Fervidobacterium* spp.'corresponds to accession number WP_011994748.1 and is found
807 in *Fervidobacterium nodosum* Rt17-B1 (NC_009718.1), *Fervidobacterium pennivorans*
808 DSM 9078 (NC_017095.1), *Fervidobacterium islandicum* (NZ_CP014334.1),
809 *Fervidobacterium gondwane*nse DSM 13020 (FRDJ01), *Petrotoga olearia* (PNR98053)
810 and *Coprothermobacter proteolyticus* (PXJB01). The protein labelled 'Bacteria inc.
811 *Mahella australiensis* MAV1, *Pseudothermotoga elfii*' corresponds to accession number
812 WP_013782344.1 and is also found in *Clostridium* sp. SYSU GA15002T
813 (NZ_CP040924.1), *Thermoanaerobacter thermocopriae* JCM 7501 (NZ_KI912455.1),
814 *Pseudothermotoga elfii* and MAV1 from *Mahella australiensis.* The tree was rooted by

mid-point rooting and visualized using iTOL (Letunic and Bork, 2019). Tree scale, substitutions per site.

Fig. 5. Electron micrographs of the induced virus and vesicles, stained with 2% uranyl acetate. Panel a. The TAV1 virus particle, which shows a typical siphovirus morphology. **Panel b.** Vesicles and toga fragments produced by *Thermosipho africanus* H17ap60334. **Panel c.** Vesicles produced by *Petrotoga* sp. 8T1HF07.NaAc.6.1, some of which are attached to a flagellum. **Panel d.** Sunflower-like structures inside *Petrotoga* sp. 8T1HF07.NaAc.6.1 cells. The structures are highlighted by arrows.

Supporting Fig. S1. Panel A. Gene-sharing network of proviruses calculated in VContact2. The network is based on shared protein clusters between viral genomes. Only proviruses at most three nodes away from MPV1 and P8T1HF07V1 are shown. The Thermotogota proviruses are colored in red, viruses from Firmicutes are blue and viruses infecting other taxa are colored orange. The quality scores calculated by ClusterOne are 0.94 (p=0.00004) for the Group 1 cluster and 0.83 (p=0.006) for the Group-2 cluster. **Panel B. Placement of proviruses on the phylogenetic tree of Thermotogota genomes reconstructed from 74 single copy protein-coding genes**. Closely related genomes (distance > 0.1), monophyletic genomes from the same genus, and clades consisting of only metagenome assembled genomes were collapsed. Identified proviruses are indicated next to their respective host genera. The tree was visualized in iTOL (Letunic and Bork, 2019). Tree scale, substitutions per site.

Supporting Fig. S2. **Placement of complete Thermotogota and Firmicutes proviruses on the viral proteomic tree.** The viral proteomic tree is from ViPTree v. 1.9 (Nishimura *et al.*, 2017), and only the relevant region of the tree is shown. The Thermotogota and Firmicutes proviruses are labeled with red stars. Taxonomy of the related viruses and their hosts is indicated as color bars next to a terminal leaf on the tree.

Supporting Fig. S3. Maximum likelihood trees of three families of terminase large subunit genes. The phylogenetic trees displayed were constructed using RAxML as implemented in Geneious v. 10 with a WAG+G substitution model and 100 bootstrap replicates. The trees should be considered unrooted. Bootstrap support > 70% is shown on branches as circles, with the size corresponding to the strength of support. Taxonomic labels of Thermotogota with proviruses are shown in red bold font, with the provirus name listed after the host name. Thermotogota homologs from genomes with no detected provirus are listed in bold font. Numbers in front of each taxon name represent database accession numbers. The tree was visualized in iTOL and rooted by midpoint rooting and should be considered unrooted (Letunic and Bork, 2019).

Supporting Fig. S4. Overview of CRISPR spacer sequences from Thermotoga isolates from the Troll oil reservoir mapped on to the TAV1 genome. Alignment position of each CRISPR spacer is indicated as black bars. Mapping and visualization was performed in Geneious v. 10 and maximum of one mismatch was allowed.

Supporting Fig. S5. Comparison of the three complete and two incomplete Thermotogota Group 2 provirus sequences. Virus name and the genus the host belongs to is indicated. The regions with significant pairwise BLASTX similarity scores are connected, red indicates that sequence is in the same direction while blue indicates that the similar sequences are on opposite strands. The predicted ORFs are color-coded based on their function and should be considered approximate, because it relies only on gene annotations. Selected 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), dimethyladenosine trf: dimethyladenosine transferase, 2Fe-2S bp: 2Fe-2S binding p rotein, MFS transporter: multi facilitator superfamily transporter. The figure was produced using genoPlotR (Guy *et al.*, 2010).

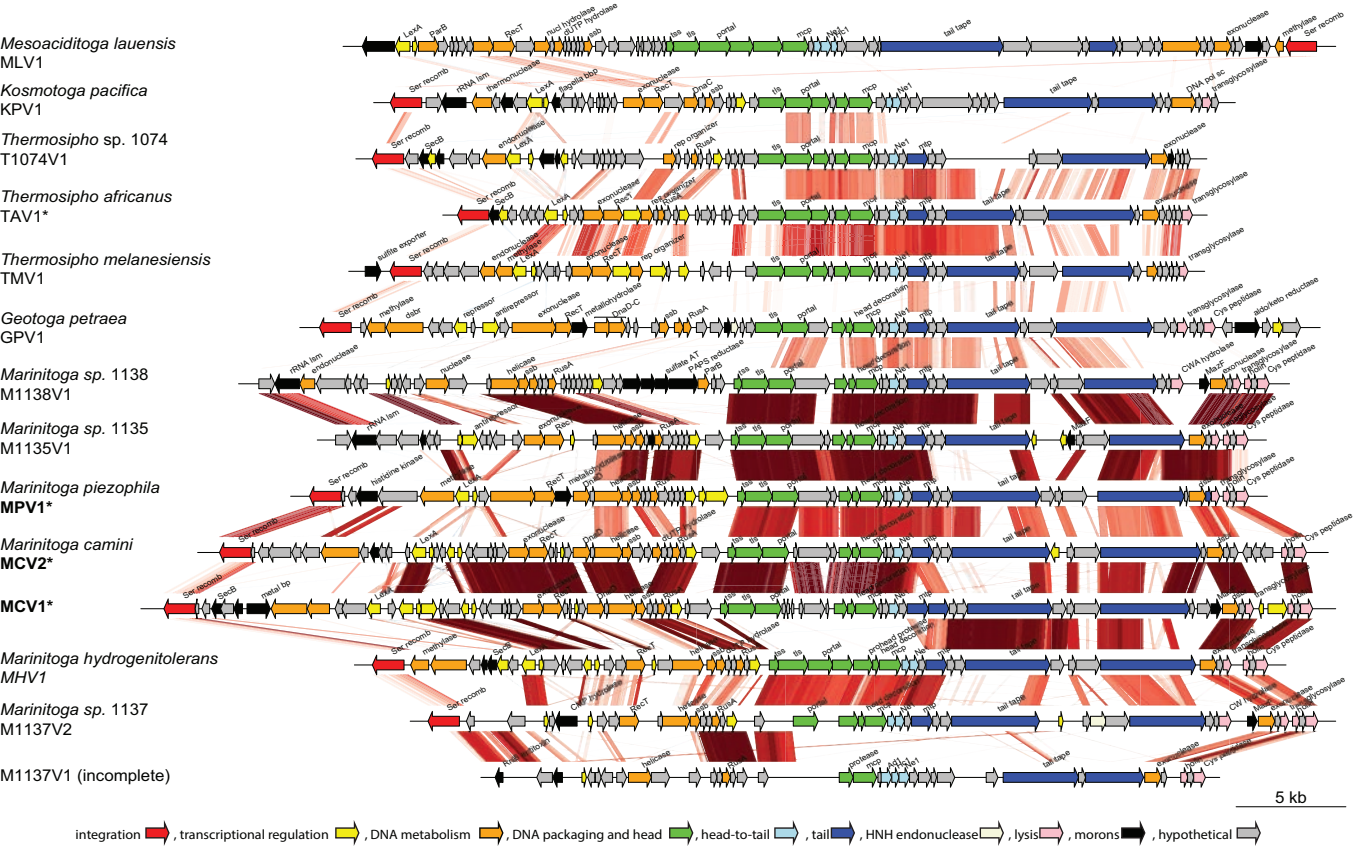
Supporting Fig. S6. Maximum likelihood trees of three potential virus receptor genes. Panel A: Competence protein ComEA, Panel B: oxaloacetate decarboxylase and Panel C: ATP-binding cassette, subfamily B. Bootstrap support > 70% is shown on branches as circles, with the size corresponding to the strength of support. The names of Thermotogota taxa that contain Group 2 proviruses are displayed in red font and Firmicutes with Group 2-like proviruses are displayed in blue font. Clades containing sequences from the same genus are collapsed into wedges. The trees were rooted using midpoint rooting, and should be considered unrooted. The trees were visualized in iTOL (Letunic and Bork, 2019).

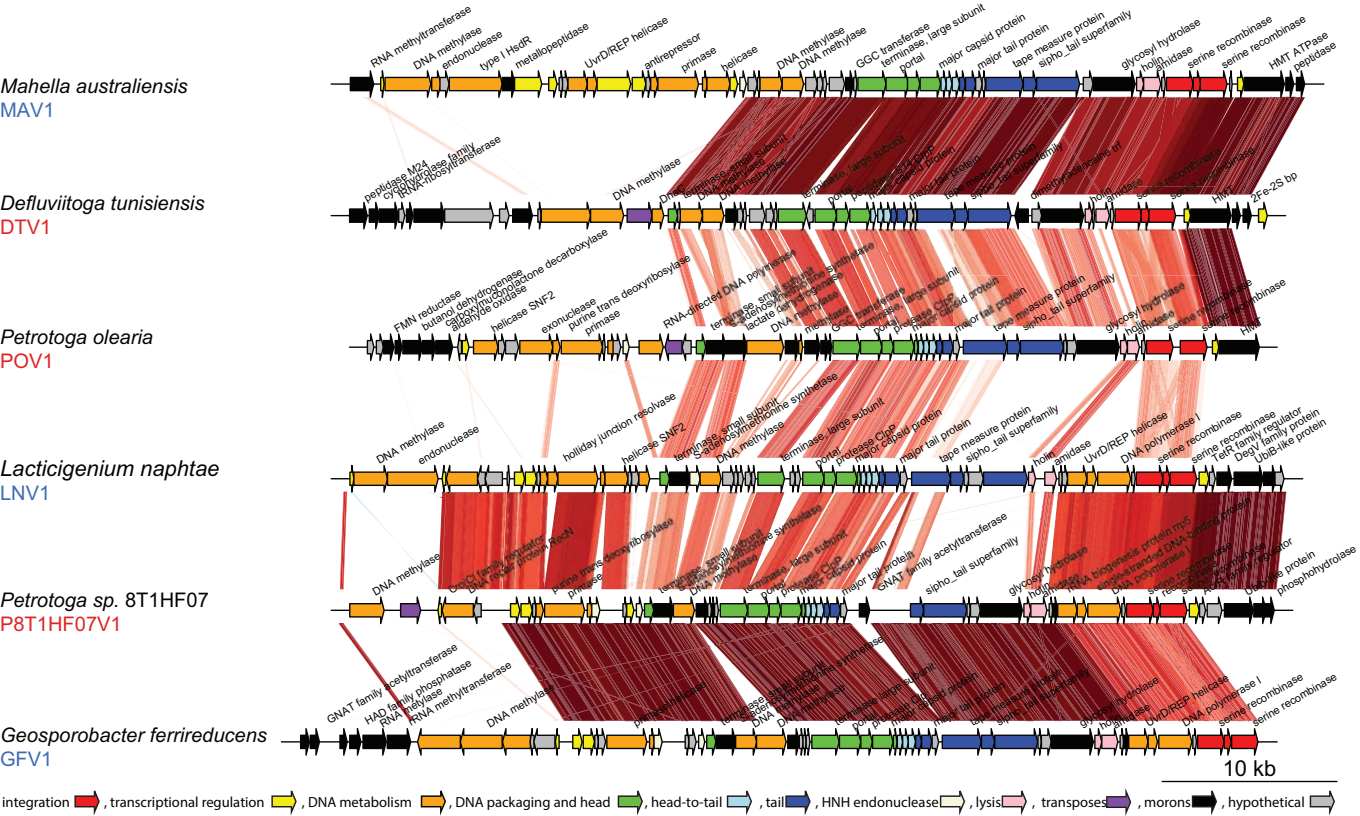
Supporting Fig. S7. Overview of the alignment of the ABC transporter ATP-binding protein in Thermotogota and Firmicutes genomes with Group 2 proviruses. Sites 100% conserved in all sequences sites are highlighted in color, while variable sites are shown in grey. The average pairwise sequence identity for the full alignment was 63% over the full alignment and 71% for the predicted C-terminal extracellular domain. Transmembrane regions, predicted using the TMHMM Server v. 2.0, are shown in red above the alignment (http://www.cbs.dtu.dk/services/TMHMM-2.0/).

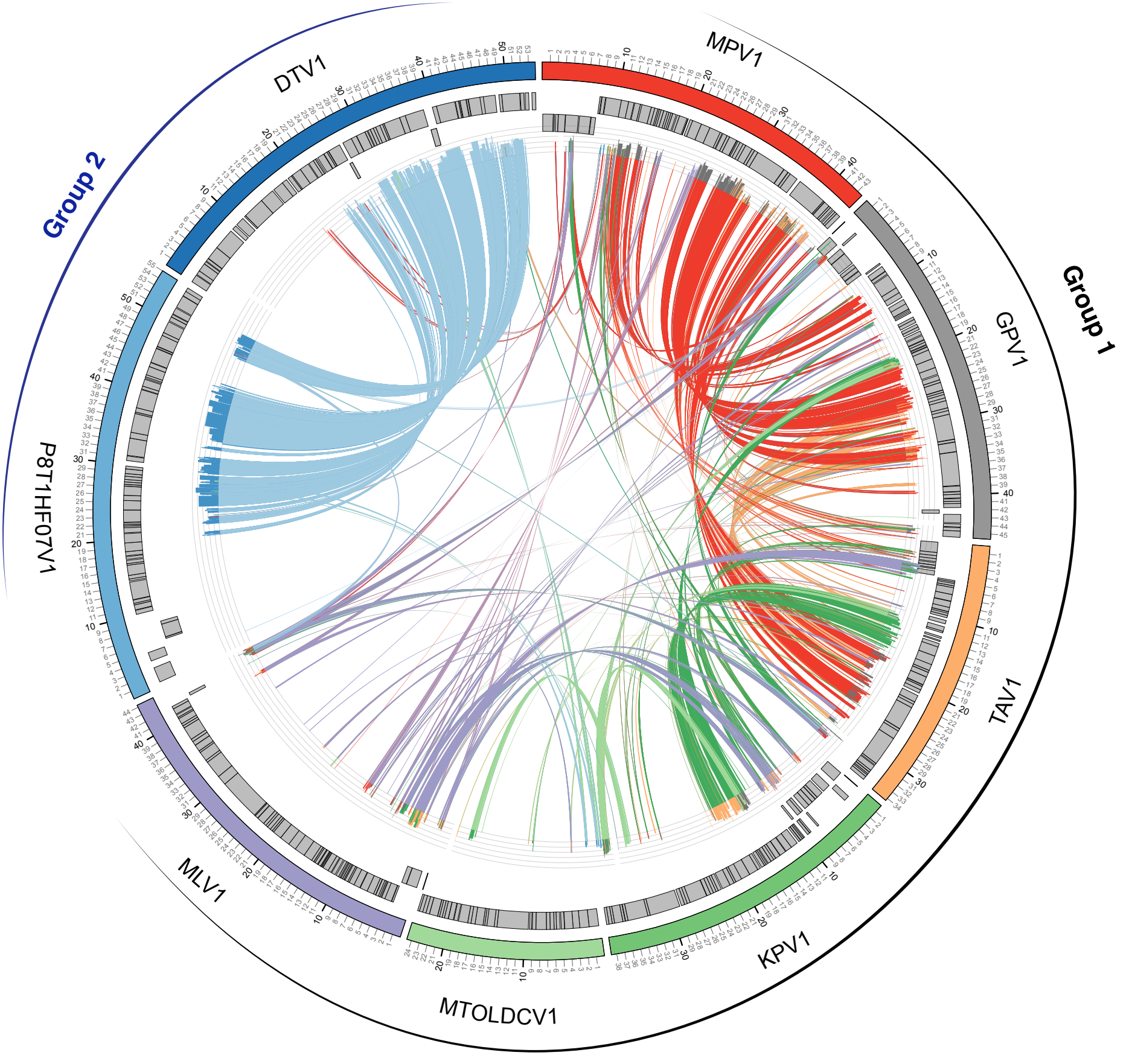
Supporting Fig. S8. Maximum likelihood tree of the moron gene annotated as a cadmium transporter. The trees was constructed using RAxML as implemented in Geneious v. 10 with a WAG + G substitution model and 100 bootstrap replicates. Bootstrap support > 70% is shown on branches as circles, with the size corresponding to the strength of support. Taxon names of Thermotogota with a provirus are given in red and taxon name of Firmicutes with a Group 2-like provirus are given in blue. Provirus name is also indicated. Thermotogota homologs from genomes with no detected provirus, or where the homolog is found outside the provirus region, are given in bold font. Database accession numbers are shown in front of taxonomic names. The tree was rooted by midpoint rooting and visualized in iTOL (Letunic and Bork, 2019).

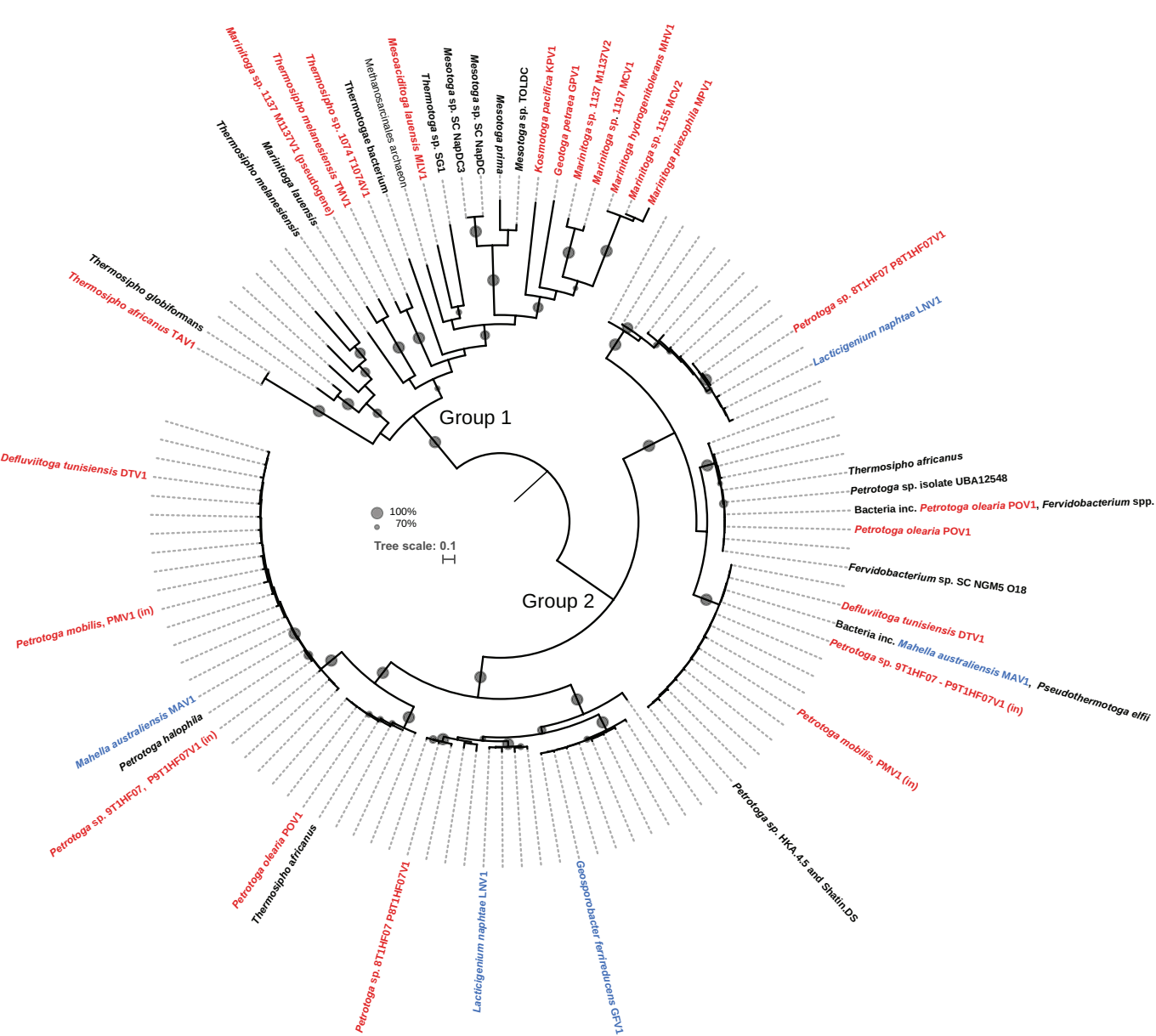
Supporting Fig. S9. Maximum likelihood trees of provirus genes with homologs in Thermotogota genomes with no detectable provirus. Database accession numbers are shown before taxonomic names. The trees were constructed using FastTree with the JTT+G (Price *et al.*, 2010) and drawn using FigTree v.1.4.4

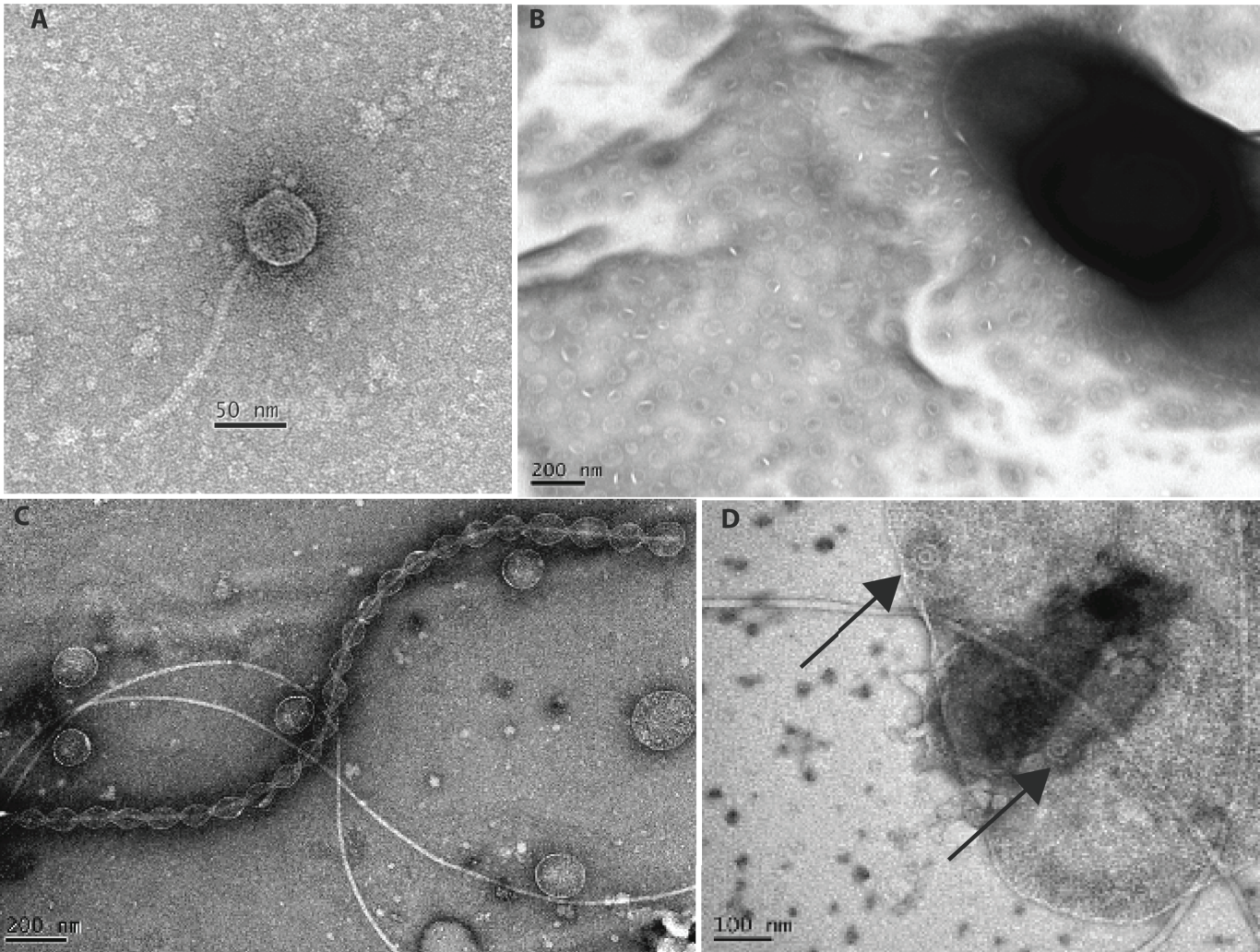
(https://github.com/rambaut/figtree/). Numbers on branches refer to FastTree support values > 0.7. Proteins from Thermotogota proviruses are given in red and proteins from the Group 2-like Firmicutes proviruses are in blue. Clades of sequences from the same genus were collapsed and shown as triangles.


Mesoaciditoga lauensis
MLV1
Kosmotoga pacifica
KPV1
Thermosipho sp. 1074
T1074V1
Thermosipho africanus
TAV1*
Thermosipho melanesiensis
TMV1
Geotoga petraea
GPV1
Marinitoga sp. 1138
M1138V1
Marinitoga sp. 1135
M1135V1
Marinitoga piezophila
MPV1*
Marinitoga camini
MCV2*
MCV1*
Marinitoga hydrogenitolerans
MHV1
Marinitoga sp. 1137
M1137V2
M1137V1 (incomplete)
5 kb
integration , transcriptional regulation , DNA metabolism , DNA packaging and head , head-to-tail , tail , HNH endonuclease , lysis , morons , hypothetical


Mahella australiensis
MAV1
Defluviitoga tunisiensis
DTV1
Petrotoga olearia
POV1
Lacticigenium naphtae
LNV1
Petrotoga sp. 8T1HF07
P8T1HF07V1
Geosporobacter ferrireducens
GFV1
10 kb
integration , transcriptional regulation , DNA metabolism , DNA packaging and head , head-to-tail , tail , HNH endonuclease , lysis , transposes , morons , hypothetical


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


Marinitoga sp. 1137 M113TV1 (pseudogene)
Thermosipho sp. 1074 T107AV1
Thermosipho melanesiensis TMV1
Mesaciditoga lauensis MLV1
Thermotoga sp. SG1
Mesotoga sp. SC NapDC3
Mesotoga prima
Mesotoga sp. TOLDC
Kosmotoga pacifica KPV1
Geotoga petraea GPV1
Marinitoga sp. 1137 M113TV2
Marinitoga sp. 1197 MCV1
Marinitoga hydrogeniotolerans MHV1
Marinitoga sp. 1155 MCV2
Marinitoga piezophila MPV1
Thermosipho melanesiensis
Marinitoga lauensis
Thermotogae bacterium
Methanobacteriales archaeon
Thermosipho globiformans
Thermosipho africanus TAV1
Petrotoga sp. 8T1HF07 P8T1HF07V1
Lacticigenium naphtae LNV1
Group 1
100%
70%
Tree scale: 0.1
Group 2
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