

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

3

4 Running title: Viruses and Proviruses of Thermotogota.

5

6 Thomas H. A. Haverkamp^{a#}, Julien Lossouarn^b, Olga Zhaxybayeva^c, Jie Lyu^d, Nadège
7 Bienvenu^d, Claire Geslin^d and Camilla L. Nesbø^{e,f}

8

9 ^aCentre for Ecological and Evolutionary Synthesis, Department of Biosciences,
10 University of Oslo, Norway.

11 ^bUniversité Paris-Saclay, INRAE, AgroParisTech, Micalis Institute, 78350,
12 Jouy-en-Josas, France

13 ^cDepartment of Biological Sciences, Dartmouth College, Hanover, NH, USA.

14 ^dUniversité Brest, CNRS, IFREMER, Laboratoire de Microbiologie des Environnements
15 Extrêmes, F-29280 Plouzané, France

16 ^eDepartment of Biological Sciences, University of Alberta, Edmonton, AB, Canada

17 ^fDepartment of Chemical Engineering and Applied Chemistry,
18 University of Toronto, Toronto, ON, Canada, M5S 3E5

19

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

22

23

24 **ABSTRACT**

25 Phylogenomic analyses of bacteria from the phylum Thermotogota have shown extensive
26 lateral gene transfer (LGT) with distantly related organisms, particularly with Firmicutes.
27 One likely mechanism of such DNA transfer is viruses. However, to date only three
28 temperate viruses have been characterized in this phylum, all infecting bacteria from the
29 *Marinitoga* genus. Here we report 17 proviruses integrated into genomes of eight
30 Thermotogota genera and induce viral particle production from one of the proviruses. The
31 proviruses fall into two groups based on sequence similarity, gene synteny and taxonomic
32 classification. Proviruses of one group are found in six genera and are similar to the
33 previously identified *Marinitoga* viruses, while proviruses from the second group are
34 only distantly related to the proviruses of the first group, have different genome
35 organization and are found in only two genera. Both groups are closely related to
36 Firmicutes in genomic and phylogenetic analyses, and one of the groups show evidence
37 of very recent LGT and are therefore likely capable of infecting cells from both phyla.
38 We conjecture that viruses are responsible for a large portion of the observed gene flow
39 between Firmicutes and Thermotogota.

40

41 **Introduction**

42 The phylum Thermotogota comprises anaerobic fermentative bacteria, most of which are
43 thermophiles [1] . They are common in subsurface environments such as marine vents,
44 terrestrial hot springs and deep subsurface oil reservoirs [2–5]. On phylogenetic trees of
45 16S rRNA gene, Thermotogota are usually a deep branching bacterial lineage, while
46 ribosomal proteins and other markers do not always agree with that placement [6, 7].
47 Such discrepancies are likely due to lateral gene transfer (LGT), which has been an
48 important force shaping the genomes of Thermotogota, with Firmicutes and Archaea
49 being their most notable gene transfer partners [1, 7, 8]. The LGT between Firmicutes
50 and Thermotogota is so extensive that the two phyla have been suggested to be linked by
51 “highways of gene sharing” [7]. However, how these inter-phylum gene-sharing events
52 occur is still unclear.

53 The subsurface constitutes the largest biosphere on Earth and is estimated to
54 contain ~70% of all cells [9]. Viruses are likely to be particularly important in subsurface
55 environments, since 97% of all viruses on earth being found in soil and sediments [10,
56 11]. Moreover, although both prokaryotic cell and virus numbers decrease with depth, the
57 virus-to-cell ratio increases with depth [10, 12, 13]. Phylogeographic studies of
58 hyperthermophilic *Thermotoga* and mesophilic *Mesotoga* have revealed genetic
59 interaction between geographically distant populations, particularly among the
60 hyperthermophilic *Thermotoga* [3, 5]. Viruses are one potential source of such long-
61 distance dispersal of genetic material [14], especially for anaerobic organisms where
62 surface dispersal is problematic.

63 Although viruses are likely candidates for transferring DNA both within and
64 between species, only three temperate siphoviruses (MCV1, MCV2, and MPV1), all
65 infecting one Thermotogota genus, *Marinitoga*, have been described [15, 16]. MCV1 and
66 MCV2 infect *Marinitoga camini* strains isolated from deep-sea hydrothermal vents [16].
67 MPV1 infects the deep-sea marine vent bacterium *Marinitoga piezophila*, where it is
68 highjacked by a plasmid co-occurring in the same host, illustrating the potential route of
69 gene mobilization in these ecosystems [15]. The three viruses are found as proviruses in
70 their host genomes and show similar genomic organization and virion morphology.
71 Phylogenetic and protein sequence-similarity analyses of the viral ORFs revealed that
72 they often group either with Firmicutes or Firmicutes' viruses, which suggests that
73 viruses infecting members of Firmicutes and Thermotogota phyla share a common gene
74 pool [15, 16].

75 Here we report 17 additional proviruses in Thermotogota genomes from eight
76 Thermotogota genera, and a successful induction of one of these proviruses. The
77 identified proviruses fall into two distinct groups. Both groups are closely related to
78 Firmicutes viruses, and the proviruses from one of these groups are likely able to infect
79 cells from both phyla. We hypothesize that membrane transport proteins, such as ABC
80 transporters, serve as receptors for Thermotogota viruses. We propose a mechanism that
81 could account for the highways of gene sharing observed between Thermotogota and
82 Firmicutes, where LGT of viral genes encoding transmembrane proteins may make the
83 host vulnerable to new viruses.

84

85 **Material and Methods**

86 **Prediction and taxonomic classification of proviruses and functional annotation of
87 their ORFs**

88 One hundred eleven Thermotogota genomes were downloaded from either Genbank or
89 IMG [17] prior to June 2018. For draft genomes, the contigs were combined into
90 ‘artificially closed’ genome using the “union” command from the EMBOSS package
91 (version 6.6.0) [18]. Each genome was screened for the presence of proviruses using the
92 Prophinder web server [19], PHAST web server [20], and PhiSpy (version 2.3) [21]
93 between October 2014 and July 2018 (**Supplementary Table S1**). For artificially closed
94 genomes, the proviral regions that crossed contig borders were discarded. Putative
95 provirus regions were inspected to identify the most likely provirus sequence by 1:
96 looking at annotations, 2: identifying possible flanking tRNA genes and 3: comparing the
97 region to genomes from the same genus and defining the boundaries, to ensure that
98 flanking genomic regions present in closely related genomes without provirus were not
99 included. Proviruses were considered complete if they contained modules for lysogeny,
100 replication, packaging, head/tail morphogenesis and lysis. If one of these modules were
101 missing, the provirus was scored as incomplete.

102 To see if there were close relatives of the predicted proviruses in other genomes,
103 provirus ORFs were used as queries in BLASTX (ver.2.2.26) [22] searches of the NCBI
104 non-redundant (nr) database [23] (accessed between July 2018 and March 2020). When
105 homologs for multiple ORFs from the predicted provirus were found in the same distantly
106 related subject genome (usually a Firmicutes genome), the identified genome was
107 downloaded and aligned to the Thermotogota genome carrying the provirus using
108 Progressive Mauve [24]. This resulted in the identification of a local alignment covering

109 similar proviruses in two otherwise distantly related genomes. The aligned region was
110 used to determine the boundaries of the provirus, limiting the provirus ends to the ends of
111 the alignment.

112 Proviral-ORF annotations were obtained from their respective Genbank entries
113 and supplemented by results from BLASTP searches [22] of the *nr* database with an
114 expected-value cutoff of 10^{-1} , and from HHpred searches [25] of the PDB database [26]
115 with a probability cutoff of 99%. In addition, recombinase- and terminase-encoding
116 ORFs were annotated using InterProScan [26], as implemented in Geneious v.10
117 (Biomatters Ltd.).

118 The sequences of the predicted proviruses were compared to each other using
119 BLASTN and TBLASTX (ver.2.2.26) [22] and visualized using genoPlotR [27] and
120 Circos [28]. Taxonomic classification of the provirus genomes was carried out using
121 searches of NCBI's viral RefSeq database (v. 94) as implemented in VContact2 using
122 Diamond [29] to identify viral protein clusters and ClusterONE [30] to obtain virus
123 clusters [31, 32]. Taxonomic classification was also assessed with Virfam [33], VIRIDIC
124 [34] and VIPTree [35]. Morphological classification was obtained with Virfam [33].

125

126 **Inference of potential host range of the putative Thermotogota viruses.** A database
127 containing all proteins from 59 Thermotogota genomes without identified proviruses was
128 constructed in Geneious v.10. Translated Thermotogota provirus proteins were used as
129 queries in BLASTP searches of this database. The provirus genes were scored as present
130 in the Thermotogota genomes if the query protein had a match with $> 50\%$ amino acid
131 identity and $> 60\%$ coverage.

132 CRPISPR spacer sequences from 90 *Thermotogota* genomes were obtained from
133 IMG [17]. The spacers were mapped to the provirus genomes in Geneious v.10, allowing
134 upto 10% nucleotide mismatches.

135

136 **Phylogenetic analyses of provirus genes and candidate receptor genes.** Homologs of
137 provirus genes selected for phylogenetic analysis were obtained by searching each
138 translated proviral gene against *nr* database (accessed between December 2019 and
139 March 2020), as well as a local database of all the *Thermotogota* virus proteins identified,
140 using BLASTP (version 2.2.26), with E-value cutoff of 10^{-1} . The 20 top-scoring matches
141 from each database were retrieved and aligned using MAFFT v. 7.450 with the G-INS-I
142 option [36]. Identical sequences and highly similar sequences from the same genus were
143 removed. Alignment positions with > 50% gaps were trimmed. Phylogenetic trees were
144 reconstructed using RAxML [37] with WAG+G substitution model with four rate
145 categories and 100 bootstrap replicates, as implemented in Geneious v.10.

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

154 Phylogenetic analysis of single copy gene in Thermotogota genomes available in
155 Genbank (accessed May 27 2020) were done using the GToTree pipeline [38] with the
156 Bacterial hmm-set of 74 target genes. The resulting alignment was imported into
157 Geneious Prime 2020.1.2 where sites with more than 50% gaps were removed, giving an
158 alignment of 11,003 amino acid positions. The phylogenetic tree was reconstructed using
159 FastTree with the JTT model with optimized Gamma20 likelihood [39].

160

161 **Virus induction and electron microscopy.** *T. africanus* H17ap6033 and two *Petrotoga*
162 isolates, *P.olearia* and *Petrotoga* sp. 8T1HF07.NaAc.6.1, were cultivated in a modified
163 Ravot medium as previously described [15] at 65°C and 55°C, respectively. Attempts
164 were made to increase the viral production of the strains by using mitomycin C, as
165 reported previously [15, 16]. A final concentration of 5 µg/mL of mitomycin C was
166 added to 300 mL bacterial culture at early to mid-log growth phase. After 3 hours of
167 incubation with mitomycin C, cultures were centrifuged at 7500 rpm and 4°C for 15 min,
168 and supernatants were ultracentrifuged at 37 000 rpm (~100 000 g) and 10°C for 1h
169 (Beckman Optima LE-80 K; rotor 70.1.Ti). Pellets were resuspended in 100µL of buffer
170 (10 mM Tris-HCL, 100 mM NaCl, 5 mM CaCl₂, 20 mM MgCl₂) and suspensions were
171 prepared for negative staining electron microscopy as previously described [40]. Briefly,
172 5 µL of the suspensions were directly spotted onto a Formvar carbon coated copper grid.
173 Putative virus-like particles were allowed to adsorb to the carbon layer for 2 min and
174 excess of liquid was removed. 5 µL of a staining uranyl acetate solution (2%) was then
175 spotted to the grid for 45 s and excess of liquid was removed again. The grid was imaged
176 at 120 kV in a JEOL JEM 100 CXIIVR transmission electron microscope.

177

178 **Results**

179

180 **Newly identified Thermotogota proviruses come from two distinct viral**
181 **lineages.** Analysis of 111 Thermotogota genomes identified 20 proviruses, including the
182 three already characterized viruses from *Marinitoga* [15, 16] and four likely partial
183 proviruses (**Supplementary Table S1** and **Supplementary Table S2**). One of the 20
184 proviruses is present with 100% nucleotide sequence identity in all six available
185 *Thermosiphon melanesiensis* genomes [41], and therefore is counted as just one novel
186 provirus. An additional provirus (MLaV1) was reported in a *Marinitoga lauensis* genome
187 after we completed the screening [42]. Due to its similarity to proviruses identified in
188 other *Marinitoga* genomes, it was not included in our further analyses.
189
190 The predicted proviruses can be divided into two distinct groups, hereafter
191 denoted as Group 1 (15 proviruses: 13 complete and 2 incomplete) and Group 2 (5
192 proviruses: 3 complete and 2 incomplete). First, the genome organization differs between
193 the proviruses in two groups (**Fig. 1** and **Fig. 2**). Second, the genes within each group are
194 more similar than the genes between groups (**Fig. 3**, **Supplementary Table S2**). Third,
195 the two groups form separate clusters in the VContact2 network (**Supplementary Fig.**
196 **S1, panel A**). Finally, the two groups show up as different clades on the Viral Proteomic
197 Tree (**Supplementary Fig. S2**). Group 1 proviruses are found in the genera *Marinitoga*,
198 *Thermosiphon*, *Kosmotoga*, *Mesotoga*, *Geotoga* and *Mesoaciditoga*, while Group 2
199 proviruses are limited to the genera *Petrotoga* and *Defluviitoga* (**Supplementary Fig. S1**,
panel B). However, presence of 29 protein families shared between the two groups

200 (Supplementary Table S3) suggests that LGT may occur between the viruses of the two
201 groups.

202

203 **Classification and genomic features of Group 1 proviruses.**

204 All complete Group 1 proviruses are likely to encode siphoviruses based on their head,
205 neck and tail gene sequences [33], and the morphology observed for the earlier
206 characterized MPV1, MCV1, MCV2 viruses and the TAV1 virus induced in the current
207 study (see below).

208 None of the proviruses have significant nucleotide identity with viral genomes in
209 the NCBI nr/nt database. Following taxonomic criteria, where viruses with > 50-70%
210 nucleotide identity over the full genome belong to the same genus and viruses with >
211 95% nucleotide identity belong to the same species [43, 44], the complete proviruses are
212 assigned to 13 new species and at least 11 new genera (Supplementary Table S4). The
213 sequence similarity suggests that the closest relatives of the Group 1 proviruses are
214 Firmicutes' viruses (Supplementary Fig. S1, panel A), since 15% of the provirus genes
215 have Firmicutes as the top-scoring match, if members of the proviruses' host genus are
216 excluded (Supplementary Table S5).

217 The proviruses have the same modular structure as genomes of the earlier
218 described MPV1, MCV1 and MCV2 viruses [15, 16] (Fig. 1). The 5' module contains
219 genes involved in lysogeny and is encoded on the opposite strand compared to the rest of
220 the virus genes. The lysogeny module is followed by modules for replication, packaging,
221 morphogenesis and host lysis. Similar to the described *Marinitoga* viruses [15, 16], the

222 gene content of lysogeny module of all examined proviruses is very variable, with only
223 the recombinase gene conserved (**Fig. 1**).

224 All Group 1 proviruses are inserted next to a tRNA gene. Eight of them (the
225 *Marinitoga* proviruses MPV1, MCV1, MCV2, MHV1 and M1137V2; the *Kosmotoga*
226 *pacifica* provirus KPV1, *Geotoga petrae* provirus GPV1, and *Mesoaciditoga lauensis*
227 provirus MLV1), are inserted next to the tRNA-Glu gene, and carry similar site-specific
228 DNA serine recombinases (KEGG Orthology; K06400, homologs of Marpi_0291 in
229 MPV1 from *M. piezophila*) (**Fig. 4**). The *Thermosiphlo* proviruses TMV1 and T1074V1
230 carry more distant homologs of this recombinase (**Fig. 4**) and are inserted next to the
231 tRNA-Phe gene. The most divergent homolog of the serine recombinase is present in
232 TAV1, which is inserted next to the tRNA-Pro gene. MTOLDCV1 is also inserted next to
233 a tRNA-Pro gene, but it is located at the end of a contig in an incomplete single cell
234 genome and its 5' end (where recombinase would be found) is missing (the *Mesotoga* sp.
235 SC_TOLDC recombinase included in **Fig. 4** is located on a separate contig). We
236 hypothesize that these recombinases are integrases that specifically recognize the tRNA
237 genes next to which the provirus is inserted (i.e., the tRNA-Glu, tRNA-Phe, or tRNA-Pro
238 genes). The remaining three proviruses (M1135V1, M1138V1, and M1137V1) are
239 inserted next to tRNA-Cys gene and may also use similar integration mechanism, but
240 these proviruses do not have the detectable serine recombinase homologs. M1135V1 and
241 M1138V1 have homologous ORFs of unknown function in the recombinase gene
242 position (**Fig. 1**), which may or may not provide this function.

243 Another typical viral protein that shows variation across the Group 1 proviruses is
244 the large subunit of the terminase protein involved in the packaging of viral DNA into the

245 virus particle [45]. The proviruses carry three types of these proteins (BLASTP E-value
246 cut-off < 0.01, identity < 25%, **Supplementary Fig. S3**). The first type, exemplified by
247 the protein in MPV1 (Marpi_0320), contains a PBSX family domain and has homologs in
248 MCV1 and MCV2 and the proviruses in *Marinitoga* sp. 1135, *Marinitoga* sp. 1138,
249 *Thermosiphon* sp. 1074 and *M. lauensis*. The second, exemplified by the terminase in
250 TAV1 (H17ap60334_04902), contains a ‘Terminase_lsu_T4-like’ domain and has
251 homologs in the proviruses from *K. pacifica*, *T. melanesiensis*, *Mesotoga* sp. TolDC and
252 *G. petrae*. The third type is found in the proviruses in *M. hydrogenitolerans* and
253 *Marinitoga* sp. 1137 (BUA62_RS02495, LN42_01905 and LN42_00550). These
254 terminases also contain a ‘Terminase_lsu_T4-like’ domain and are distant homologs of
255 the second terminase type.

256 In addition to the recombinase and terminase, other typical viral proteins such as
257 tail tape measure, capsid and portal proteins were identified, but did not always show
258 detectable similarity among the proviruses (**Fig. 1**). Two transcription regulators
259 (Marpi_0297 and Marpi_0298 in MPV1), a DNA repair exonuclease (Marpi_0340 in
260 MPV1) and a single stranded DNA-binding protein (Marpi_0306 in MPV1), show
261 relatively high (32-100%) identity across most Group 1 proviruses (**Fig. 1**,
262 **Supplementary Table S4**). Genes encoding two hypothetical proteins (homologs of
263 Marpi_0299 and Marpi_0338 in MPV1) are shared among 10 of the proviruses (36-96%
264 identity), suggesting these genes may provide important viral functions.

265

266 **Broad host range of Group 1 proviruses.**

267 Detection of Group 1 proviruses in the genera *Marinitoga*, *Thermosiphho*, *Kosmotoga*,
268 *Mesotoga*, *Geotoga* and *Mesoaciditoga* (**Supplementary Table S1**), suggests that the
269 Group 1 viruses are widespread among Thermotogota, particularly among organisms
270 inhabiting hydrothermal vents. Such wide distribution and relatively high sequence
271 identity among the proviral genomes (**Fig. 1, Supplementary Table S4**) suggest that the
272 Group 1 temperate viruses might have broad host ranges. Experiments showing that
273 MPV1 from *M. piezophila* can infect and transfer a plasmid to a *Thermosiphho* isolate is
274 consistent with this hypothesis [15].

275 Further support comes from mapping of CRISPR spacer sequences from 90
276 Thermotogota genomes to the Group 1 proviruses. Five of the 17 proviruses matched
277 CRISPR spacers in the genomes from a different genus (**Table 1**). For example, the
278 *Thermosiphho* provirus TAV1 had 35 matches to spacers in the genomes of
279 *Pseudothermotoga* and *Thermotoga* spp. Anecdotal evidence corroborates an ability of
280 TAV1 to infect *Thermotoga* spp. Back in 2005, when the sample from the Hibernia oil
281 reservoir containing TAV1 and its host, *T. africanus* H17ap60334, was being processed
282 by one of us (Camilla L. Nesbø) in the laboratory, *Thermotoga* isolates from Troll oil
283 reservoir in the North Sea, which were at the same time being transferred to fresh media,
284 experienced a mass death. Analysis of the genomes of the surviving *Thermotoga* isolates
285 [3] revealed presence of three CRISPR spacer matching the TAV1 genome
286 (**Supplementary Fig. S4**). These spacers were located in the middle of the CRISPR
287 arrays, indicating that they were not new acquisitions [46]. Therefore the only surviving
288 isolates of *Thermotoga* must have had already experienced and survived TAV1 or related
289 virus infections in the oil reservoir.

290

291 **Classification, genomic features and distribution of Group 2 proviruses.**

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

297 Following the taxonomic classifications criteria described above, the three
298 complete proviruses P8T1HF07V1, DTV1 and POV1 are assigned to three new viral
299 species (**Supplementary Table S4**). Based on the head-neck-tail module classification
300 [33], these proviruses likely encode siphoviruses of Type1 - Cluster 2. All hosts of the
301 previously described members of this *Siphoviridae* lineage belong to Firmicutes. In
302 agreement with this, similarity searches revealed that these proviruses show very high
303 similarity to proviruses of three Firmicutes genomes: *Lacticigenium naphtae* (LNV1),
304 *Geosporobacter ferrireducens* (GFV1) and *Mahella australiensis* (MAV1) (**Fig. 2**,
305 **Supplementary Table S4**).

306 The sequences and genome organization of the three complete Group 2 proviruses
307 differ considerably from that of Group 1 (**Fig. 1** and **Fig. 2**). These proviruses are also not
308 located next to tRNA genes. The 5' module encodes genes involved in virus replication
309 and transcription, and the comparative genomic analysis shows high level of diversity in
310 this region (**Fig. 2**). This module is followed by highly conserved packaging,
311 morphogenesis and lysis modules. The lysogeny module is located at the 3' end of the

312 virus. The site-specific serine recombinases carried by the Group 2 proviruses in this
313 module are distant homologs of the earlier discussed Group 1 recombinases (**Fig. 4**).

314 When comparing the Group 2 provirus genomes from Thermotogota and
315 Firmicutes, each Thermotogota provirus is more similar to a Firmicutes provirus than to
316 other Thermotogota proviruses (**Fig. 2, Supplemental Table S4**). Alignments of the two
317 Thermotogota-Firmicutes provirus pairs, P8T1HF07V1 and GFV1, and DTV1 and
318 MAV1, have 66.7 % and 56.7 % intergenomic similarity values, respectively
319 (**Supplementary Table S4**), suggesting they may be assigned to the same genus.

320 Moreover, P8T1HF07V1 has 97% nucleotide identity to the GFV1 over specific
321 subregions that encode structural genes and the genes for DNA packaging and genome
322 integration (**Fig. 3**). Similarly, in the same regions in DTV1 and MAV1 have 95-97%
323 identity. In contrast, the same regions in P8T1HF07V1 and POV1, and P8T1HF07V1 and
324 DTV1 have 53 and 75% nucleotide identity, respectively. Such similarity patterns suggest
325 that these viruses likely can infect hosts from both Thermotogota and Firmicutes phyla.

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

330

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

332 Induction assays were performed on three of the putatively lysogenized *Thermotogota*: *T.*
333 *africanus* H17ap6033 (Group 1), *Petrotoga* sp. P8T1HF07 (Group 2) and *P. olearia*
334 (Group 2). Only the provirus in *T. africanus* H17ap6033 (TAV1) was successfully

335 induced using mitomycin C. TAV1 was shown to produce viral particles with a
336 polyhedral head of ~50 nm in diameter and a flexible non-contractile tail of ~160 nm in
337 length and ~10 nm in width (**Fig. 5a**). Based on tail morphology, TAV1 was classified to
338 the order *Caudovirales* and the family *Siphoviridae*, confirming the sequence-based
339 classification. TAV1 morphology is similar to the three previously characterized
340 temperate *Marinitoga* viruses, whose virion tails were just slightly longer [15, 16]. In
341 addition to viral particles, a release of membrane vesicles or toga fragments was regularly
342 observed (**Fig. 5b**).

343 While the induction of the proviruses in *Petrotoga* sp. 8T1HF07 (P8T1HF07V1)
344 and *P. olearia* (POV1) using mitomycin C was unsuccessful, membrane vesicles of
345 various sizes and shapes (20 - 100nm) were produced by the cells, and in particular by the
346 induced *Petrotoga* sp. 8T1HF07 cells. Analysis of the supernatant of the latter culture
347 revealed similarly-sized round-shaped vesicles connected together in long chains by
348 hooking onto the flagella, like a “pearl necklace”, while free vesicles showed more
349 diversity in size and shape (**Fig. 5c**). Some “sunflower-like” structures were also
350 observed inside a remaining cell (**Fig. 5d**). It is unknown if the provirus or stressors
351 influence the production of these vesicles and structures, or if they are produced
352 spontaneously.

353

354 **A potential receptor for the Group 2 viruses**

355 Several types of structures on the surface of bacteria, such as membrane proteins,
356 flagella, pili, or carbohydrate moieties, can act as virus receptors [47]. Most siphoviruses
357 of Gram-negative bacteria, and some of Gram-positive bacteria, use proteinaceous

358 receptors for adsorption [48, 49]. If the Group 2 viruses use the same protein receptor to
359 attach to both Thermotogota and Firmicutes cells, the large phylogenetic distance
360 between these hosts offers an opportunity to identify possible membrane protein receptors
361 bioinformatically, since the receptor proteins would be expected to be conserved across
362 the genomes from both phyla. It should be noted that this approach would only identify
363 possible protein receptors, while potential shared carbohydrate receptors would not be
364 detected.

365 Four predicted membrane proteins with transmembrane helices were identified in
366 all genomes carrying a Group 2 provirus. One of these was the viral holin gene, leaving
367 three receptor candidates: a ComEA family DNA-binding protein, an oxaloacetate
368 decarboxylase beta subunit, and an ABC transporter ATP-binding protein. Phylogenetic
369 analyses revealed that the ComEA and the oxaloacetate decarboxylase homologs are
370 widely distributed among Thermotogota (**Supplementary Fig. S6**). In contrast, the ABC
371 transporter is, among the Thermotogota, restricted to *Petrotoga* and *Defluviitoga*, the two
372 genera where the Group 2 proviruses are observed (**Supplementary Fig. S6, panel C**).
373 Moreover, the phylogenetic analysis suggests the homologs in *Petrotoga* and
374 *Defluviitoga* originated from an LGT event with a Firmicute (**Supplementary Fig. S6**).
375 These proteins show particularly high amino acid sequence similarity in the C-terminal
376 domain of both Thermotogota and Firmicutes homologs, which is facing the exterior of
377 the cell and could serve as a virus target (**Supplementary Fig. S7**). Although
378 experiments are needed to demonstrate if any of these proteins functions as receptor for
379 these viruses, we suggest that the ABC-transporter ATP-binding protein is a strong
380 candidate for a Group 2 virus receptor.

381

382 **Moron genes are abundant in the identified proviruses.**

383 Many temperate viruses are known to carry moron genes, which are genes that do not
384 have a direct viral function [50, 51]. The detected proviruses of Thermotogota are no
385 exception: the Group 1 proviruses carry up to 6 morons (**Fig. 1**), while the Group 2
386 proviruses have between 4 and 13 morons (**Fig. 2**). However, it should be noted that
387 because the 5' ends were hard to define for Group 2 proviruses, some of the moron genes
388 at the 5' ends might not be part of the proviruses. Sequencing virus DNA isolated from
389 capsids will help resolve this issue in the future.

390 Among the morons are several proteins that may confer a selective advantage to
391 the host (**Fig. 1** and **Fig. 2**, **Supplementary Table S2**). For instance, M1138V1 carry two
392 genes involved in sulfur metabolism. The Group 2 proviruses encode several transporters,
393 peptidases and hydrolases, likely to be beneficial for these heterotrophic bacteria. In
394 addition, all the viruses carry several hypothetical proteins that may also have non-viral
395 functions.

396

397 **Evidence for the viruses' impact on lateral gene transfer**

398 Eight hundred seventy homologs of 106 proviral genes were detected in 54 out of 59
399 Thermotogota genomes with no detectable proviruses (**Supplementary Table S6**). It
400 should be noted that some provirus genes, e.g. the Group 1 recombinases and terminases
401 (**Fig. 4** and **Supplementary Fig. S3**), did not pass our stringent screening criteria (see
402 **Material and Methods**), thus these represent minimum estimates of matches to proviral
403 genes in these genomes. Notably, 370 of 870 were homologs of 28 moron genes,

404 suggesting that the viruses may facilitate exchange of “host” genes among Thermotogota.
405 Moron genes also had the highest number of homologs across 54 genomes, with most
406 abundant being a queuine tRNA-ribosyltransferase in the Group 2 provirus DTV1 (found
407 in 48 genomes) and an aldo/keto reductase in Group 1 provirus GPV1 (found in 41
408 genomes).

409 Among phylogenetically informative datasets, 10 proviral genes group within
410 Thermotogota and 17 group within Firmicutes, suggesting that many of the proviral genes
411 originated either in Thermotogota and Firmicutes (**Supplementary Table S6**). For
412 instance, the above-described abundant moron gene queuine tRNA-ribosyltransferase is
413 of Thermotogota origin, while the aldo/keto reductase appears to be of Firmicutes origin
414 (**Supplementary Table S6**). In the phylogeny of another moron gene, a cadmium or
415 heavy metal transporter found in Firmicutes provirus MAV1 and Thermotogota Group 2
416 proviruses DTV1 and POV1, the provirus genes group closely with Firmicutes’ homologs
417 (**Supplementary Fig. S8**). Notably, the other closely related Thermotogota homologs are
418 found in three *Fervidobacterium* and *Pseudothermotoga* genomes, genera where no
419 proviruses have yet been identified. Inspecting the genomic region surrounding these
420 genes in *Fervidobacterium* and *Pseudothermotoga*, revealed that the homolog of the
421 proviral recombinase (**Fig. 4**) is located immediately upstream of the transporter gene. No
422 other typical virus genes were observed in these regions, suggesting these genes are likely
423 remnants of proviruses. This also indicates that viruses related to Group 2 proviruses may
424 have broader host range than we presently detect.

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

428

429 **Discussion**

430 In our search for proviruses in genomes of Thermotogota, we discovered two distinct
431 groups of temperate siphoviruses that have lysogenized this bacterial phylum. These
432 proviruses may represent multiple new viral species and genera. Our analyses suggest
433 that these viruses likely have broad host range that spans at least multiple genera. We also
434 found that the identified proviruses lineages are closely related to Firmicutes' viruses.

435 One of the bioinformatically identified Group 1 proviruses (TAV1) was induced
436 and shown to produce virus particles. The provirus resides in a genome of a *T. africanus*
437 isolate from the Hibernia oil reservoir off the Canadian east coast. The analysis of
438 CRISPR spacers suggested that this virus may have a particularly wide host range, with
439 the highest number of spacer-matches in genomes from outside its genus. For instance, a
440 virus very similar to TAV1 had likely infected *Thermotoga* spp. isolates from the North
441 Sea Troll oil reservoir. Similar predatory virus pressure in geographically and
442 geologically remote subsurface environments have been observed for *Methanohalophilus*
443 isolates from reservoirs in the USA and Russia [52].

444 We were not able to induce virus production from the selected Group 2 proviruses
445 This could be due to these proviruses currently being inactive, or we may not have
446 applied the right conditions to induce the expression of these proviruses. Nevertheless,

447 the high level of sequence identity between Group 2 provirus sequences from
448 Thermotogota and Firmicutes phyla suggests that they have been active very recently.

449 Many of the genes carried by both Group 1 and Group 2 proviruses are found in
450 genomes of Thermotogota that do not have detectable proviruses. These genes often
451 group with Firmicutes or viruses that infect Firmicutes, and many can be classified as
452 morons. This suggests that both Group 1 and Group 2 viruses transfer genes within
453 Thermotogota and between Thermotogota and Firmicutes, and may serve as a major
454 mechanism for the earlier reported large amounts of lateral gene transfer between
455 Thermotogota and Firmicutes [7, 8].

456 Based on our bioinformatic and phylogenetic analyses, we propose that an ABC
457 transporter may serve as a receptor for at least some of these proviruses. ABC transporter
458 proteins are, to our knowledge, not commonly identified as bacterioviral receptors.
459 However, *Lactococcus* viruses from the siphoviral c2 group have been shown to use
460 membrane proteins Pip or YjaE, both with sequence similarity to ABC-transporter
461 domains, as secondary receptors [47, 53].

462 Intriguingly, transporters, and ABC transporters in particular, are among the most
463 frequently transferred genes both within the Thermotogota and between Thermotogota
464 and Firmicutes [7, 8, 54]. Transporter genes were also detected in the provirus genomes.
465 The possibility that transporters can function as viral receptors in the Thermotogota
466 therefore suggests that acquiring a new transporter, perhaps via a viral infection, might
467 result in the cell not only acquiring a new function but also becoming susceptible to a
468 new virus. This virus might carry another transporter gene, which can introduce yet
469 another virus, resulting in a ratchet-like process. Using transporters as receptors will

470 therefore not only provide the virus with the wide host range but could also make viruses
471 the vehicles on the highways of gene sharing observed between the *Thermotogota* and
472 Firmicutes.

473 Genes encoding proteins for membrane transport, including ABC transporters,
474 have been observed in several other viruses [55], and thus the proposed process could
475 operate widely among bacteria. This is contrary to a role commonly assigned to morons
476 where they often confer resistance to infections by other viruses [51]. Further studies and
477 experiments are needed to investigate if such ratchet processes are indeed occurring in
478 natural systems. However, regardless of the functions of the morons in the *Thermotogota*
479 proviruses, the observation of viruses potentially infecting organisms from different phyla
480 further demonstrates that viruses are key actors in the evolution of microbial diversity.

481

482 **Acknowledgements**

483 This work is supported by a Research Council of Norway award (project no.
484 180444/V40) to C.L.N., by the Sino-French LIA/PRC 1211 MicrobSea to J.L. and by the
485 Simons Foundation Investigator in Mathematical Modeling of Living Systems award
486 327936 to O.Z. Strains were obtained from the Université de Bretagne Occidentale
487 Culture Collection (UBOCC, Plouzané, France, www.univ-brest.fr/ubocc).

488

489

490 **Competing Interests:**

491 The authors declare no conflict of interest.

492

493

494 **References**

495 1. Pollo SMJ, Zhaxybayeva O, Nesbø CL. Insights into thermoadaptation and the
496 evolution of mesophily from the bacterial phylum *Thermotogae*. *Can J Microbiol*
497 2015; **61**: 655–670.

498 2. Bhandari V, Gupta RS. The Phylum Thermotogae. In: Rosenberg E, DeLong EF,
499 Lory S, Stackebrandt E, Thompson F (eds). *The Prokaryotes*. 2014. Springer Berlin
500 Heidelberg, Berlin, Heidelberg, pp 989–1015.

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

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

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

510 6. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A
511 new view of the tree of life. *Nat Microbiol* 2016; **1**: 16048.

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

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

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

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

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

526 12. Engelhardt T, Kallmeyer J, Cypionka H, Engelen B. High virus-to-cell ratios
527 indicate ongoing production of viruses in deep subsurface sediments. *ISME J* 2014;
528 **8**: 1503–1509.

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

532 14. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of
533 bacterial innovation. 2000; **405**: 6.

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

537 16. Mercier C, Lossouarn J, Nesbø CL, Haverkamp THA, Baudoux AC, Jebbar M, et al.

538 Two viruses, MCV1 and MCV2, which infect *Marinitoga* bacteria isolated from

539 deep-sea hydrothermal vents: functional and genomic analysis. *Environ Microbiol*

540 2018; **20**: 577–587.

541 17. Markowitz VM, Chen I-MA, Chu K, Szeto E, Palaniappan K, Pillay M, et al.

542 IMG/M 4 version of the integrated metagenome comparative analysis system.

543 *Nucleic Acids Res* 2014; **42**: D568–D573.

544 18. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open

545 Software Suite. *Trends Genet* 2000; **16**: 276–277.

546 19. Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. Prophinder: a

547 computational tool for prophage prediction in prokaryotic genomes. *Bioinformatics*

548 2008; **24**: 863–865.

549 20. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: A Fast Phage Search

550 Tool. *Nucleic Acids Res* 2011; **39**: W347–W352.

551 21. Akhter S, Aziz RK, Edwards RA. PhiSpy: a novel algorithm for finding prophages

552 in bacterial genomes that combines similarity- and composition-based strategies.

553 *Nucleic Acids Res* 2012; **40**: e126–e126.

554 22. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.

555 BLAST+: architecture and applications. *BMC Bioinformatics* 2009; **10**: 421.

556 23. O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al.

557 Reference sequence (RefSeq) database at NCBI: current status, taxonomic

558 expansion, and functional annotation. *Nucleic Acids Res* 2016; **44**: D733–D745.

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

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

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

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

568 28. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos:
569 An information aesthetic for comparative genomics. *Genome Res* 2009; **19**: 1639–
570 1645.

571 29. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using
572 DIAMOND. *Nat Methods* 2015; **12**: 59–60.

573 30. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-
574 protein interaction networks. *Nat Methods* 2012; **9**: 471–472.

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

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

581 33. Lopes A, Tavares P, Petit M-A, Guéris R, Zinn-Justin S. Automated classification
582 of tailed bacteriophages according to their neck organization. *BMC Genomics* 2014;
583 **15**: 1027.

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

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

588 36. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
589 Improvements in Performance and Usability. *Mol Biol Evol* 2013; **30**: 772–780.

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

592 38. Lee MD. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*
593 2019; **35**: 4162–4164.

594 39. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood
595 Trees for Large Alignments. *PLoS ONE* 2010; **5**: e9490.

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

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

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

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

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

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

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

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

616 48. Zhang Z, Yu F, Zou Y, Qiu Y, Wu A, Jiang T, et al. Phage protein receptors have
617 multiple interaction partners and high expressions. *Bioinformatics* 2020; **36**: 2975–
618 2979.

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

621 50. Cumby N, Davidson AR, Maxwell KL. The moron comes of age. *Bacteriophage*
622 2012; **2**: e23146.

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

625 52. Borton MA, Daly RA, O'Banion B, Hoyt DW, Marcus DN, Welch S, et al.
626 Comparative genomics and physiology of the genus *Methanohalophilus* , a
627 prevalent methanogen in hydraulically fractured shale. *Environ Microbiol* 2018; **20**:
628 4596–4611.

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

631 54. Nesbø CL, Nelson KE, Doolittle WF. Suppressive Subtractive Hybridization
632 Detects Extensive Genomic Diversity in *Thermotoga maritima*. *J Bacteriol* 2002;
633 **184**: 4475–4488.

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

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

638

639 **Table 1. CRISPR spacer matches to provirus genomes in Thermotogota genomes.**
640 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)
KPV1	<i>Kosmotoga olearia</i> (1)
MHV1	<i>Pseudothermotoga elfii</i> NBRC107921 (1), <i>Marinitoga</i> sp. 1154 (1)
MCV1*	<i>Marinitoga</i> sp. 1154 (2)
MCV2*	<i>Marinitoga</i> sp. 1155 (1) self-match, <i>Marinitoga</i> sp. 1154 (2)
MPV1*	<i>Marinitoga</i> sp. 1137 (1)
M1135	<i>Marinitoga</i> sp. 1155 (1)

M1137V1	<i>Pseudothermotoga elfii</i> NBRC107921 (1), <i>Marinitoga</i> sp. 1135 (2)
M1137V2	<i>Marinitoga piezophila</i> (1)
M1138	<i>Thermosiphon africanus</i> TCF52B (1), <i>Thermosiphon melanesiensis</i> (1), <i>Pseudothermotoga elfii</i> NBRC107921 (1), <i>Marinitoga</i> sp. 1137 (1)
TAV1*	<i>Thermosiphon africanus</i> H17ap60334 (3) self -match, <i>Thermosiphon africanus</i> TCF52B (2), <i>Thermosiphon africanus</i> Ob7 (1), <i>Thermosiphon melanesiensis</i> (2), <i>Thermotoga maritima</i> 2812B (1), <i>Thermotoga</i> sp. EMP (1), <i>Thermotoga</i> sp. XYL54 (3), <i>Thermotoga</i> sp. CELL2 (3), <i>Thermotoga</i> sp. TBGT1766 (3), <i>Thermotoga</i> sp. TBGT1765 (4), <i>Thermotoga</i> sp. A7A (1), <i>Thermotoga</i> sp. MC24 (2), <i>Pseudothermotoga elfii lettingae</i> (4), <i>Pseudothermotoga elfii</i> NBRC107921 (9), <i>Pseudothermotoga elfii</i> DSM9442 (2)
TMV1	<i>Thermosiphon melanesiensis</i> (1) self-match, <i>Thermosiphon africanus</i> TCF52B (1), <i>Thermosiphon africanus</i> Ob7 (1), <i>Thermotoga</i> sp. TBGT1765 (1), <i>Thermotoga</i> sp. Mc24 (1), <i>Pseudothermotoga elfii</i> NBRC107921 (1), <i>Pseudothermotoga elfii lettingae</i> (2)
T1074V1	<i>Thermosiphon affectus</i> Bl1070 (1), <i>Thermosiphon affectus</i> 1223 (3), <i>Thermosiphon affectus</i> Bl1063 (1)

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

642

643

644 **Fig. 1. Comparison of sequences from all detected Group 1 proviruses.** Provirus
645 name and the species of its host are shown to the left of the nucleotide sequence, in which

646 predicted ORFs are depicted as arrows. Proviruses that have been induced and shown to
647 produce virus particles are marked with an asterisk. The lines connect regions of adjacent
648 viruses that have TBLASTX similarity of more than 30% over 100bp. Lines are colored
649 in red or blue indicate that the matching sequences encoded in the same or opposite
650 strand, respectively. The predicted ORFs are color-coded based on their function and
651 should be considered approximate, because it relies only on gene annotations. Selected
652 gene annotations are included and abbreviated as follows. Ser recomb: serine
653 recombinase, LexA: LexA repressor, ParB: ParB-like nuclease, RecT: RecT family
654 recombinase, dUTP hydrolase: deoxyuridine 5'-triphosphate nucleotidohydrolase, ssb:
655 single stranded DNA-binding protein, tss: terminase small subunit, tls: terminase large
656 subunit, mcp: major capsid protein, tail tape: tail tape measure protein, rRNA lsm:
657 ribosomal RNA large subunit methyltransferase, flagella bbp: flagella basal-body protein,
658 DnaC: DnaC replication protein, DnaD: DnaD replication protein, DNA pol sc: DNA
659 polymerase sliding clamp, SecB: SecB protein-export protein, rep organizer :replisome
660 organizer, RusA: RusA family crossover junction endodeoxyribonuclease, Cys peptidase:
661 cysteine peptidase, sulfate AT: sulfate adenylyltransferase subunit 2, PAPS reductase:
662 phosphoadenosine phosphosulfate reductase, CW hydrolase: cell wall-associated
663 hydrolase, MazF: MazF endoribonuclease, dsbr: DNA double-strand break repair protein,
664 metal bp: metal-binding protein, CMP hydrolase: cytidine 5'-monophosphate hydrolase.
665 The figure was produced using genoPlotR [27].

666

667 **Fig. 2. Comparison of sequences from three complete Thermotogota Group 2**
668 **proviruses and their Firmicutes' homologs.** Provirus name (in red for Thermotogota

669 and blue for Firmicutes) and the species of its host are shown to the left of the nucleotide
670 sequence, in which predicted ORFs are depicted as arrows. The lines connect regions of
671 adjacent viruses that have TBLASTX similarity of more than 30% over 100bp. Lines are
672 colored in red or blue indicate that the matching sequences encoded in the same or
673 opposite strand, respectively. The predicted ORFs are color-coded based on their function
674 and should be considered approximate, because it relies only on gene annotations.
675 Selected gene annotations are included and abbreviated; HMT ATPase: heavy metal
676 translocating ATPase, FMN reductase: flavine mono nucleotide reductase, HAD family
677 phosphatase: haloacid dehalogenase superfamily of hydrolase). The figure was produced
678 using genoPlotR [27].

679

680 **Fig. 3. Comparison of representative Thermotogota proviruses.** Due to sequence
681 similarity, only one provirus per Thermotogota genus is shown. The nucleotide sequences
682 of the proviruses are arranged around the circle and color-coded. Numbers indicate
683 kilobases (kb) and grey boxes outline locations of predicted genes. Lines connecting
684 different proviral sequences represent TBLASTX matches between the proviral regions,
685 with the percent identity shown in histograms at the ends of each line. The plot was
686 created using Circos [28].

687

688 **Fig. 4. Maximum likelihood tree of recombinases found in Thermotogota proviruses**
689 **and of their homologs in Firmicutes proviruses, and Thermotogota and Firmicutes**
690 **genomes.** Host names of Thermotogota and Firmicutes proviruses are colored in red and
691 blue, respectively. The names of their proviruses are added next to the host name. Names

692 of Thermotogota homologs that either resided outside of proviral regions or come from a
693 genome without detected proviruses are shown in black. Branches without labels
694 represent Firmicutes without an identified Group 2 provirus. Homologs from incomplete
695 proviruses are labeled with “(in)”. Circles on the branches represent bootstrap support,
696 and only values above 70% are shown. Some proteins have identical amino acid
697 sequences in more than one organism. The protein labelled ‘Bacteria inc. POLV1
698 *Fervidobacterium* spp.’ corresponds to accession number WP_011994748.1 and is found
699 in *Fervidobacterium nodosum* Rt17-B1 (NC_009718.1), *Fervidobacterium pennivorans*
700 DSM 9078 (NC_017095.1), *Fervidobacterium islandicum* (NZ_CP014334.1),
701 *Fervidobacterium gondwanense* DSM 13020 (FRDJ01), *Petrotoga olearia* (PNR98053)
702 and *Coprothermobacter proteolyticus* (PXJB01). The protein labelled ‘Bacteria inc.
703 *Mahella australiensis* MAV1, *Pseudothermotoga elfii*’ corresponds to accession number
704 WP_013782344.1 and is also found in *Clostridium* sp. SYSU GA15002T
705 (NZ_CP040924.1), *Thermoanaerobacter thermocopriae* JCM 7501 (NZ_KI912455.1),
706 *Pseudothermotoga elfii* and MAV1 from *Mahella australiensis*. The tree was rooted by
707 mid-point rooting and visualized using iTOL [56]. Tree scale, substitutions per site.
708

709 **Fig. 5. Electron micrographs of the induced virus and vesicles, stained with 2%**
710 **uranyl acetate. Panel a.** The TAV1 virus particle, which shows a typical Siphoviridae
711 morphology. **Panel b.** Vesicles and toga fragments produced by *Thermosiphon africanus*
712 H17ap60334. **Panel c.** Vesicles produced by *Petrotoga* sp. 8T1HF07.NaAc.6.1, some of
713 which are attached to a flagellum. **Panel d.** Sunflower-like structures inside *Petrotoga* sp.
714 8T1HF07.NaAc.6.1 cells. The structures are highlighted by arrows.

715

716 **Supplementary Figures:**

717

718 **Supplementary Fig. S1. Panel A. Gene-sharing network of proviruses calculated in**

719 **VContact2.** The network is based on shared protein clusters between viral genomes.

720 Only proviruses at most three nodes away from MPV1 and P8T1HF07V1 are shown. The

721 Thermotogota proviruses are colored in red, viruses from Firmicutes are blue and viruses

722 infecting other taxa are colored orange. The quality scores calculated by ClusterOne are

723 0.94 (p=0.00004) for the Group 1 cluster and 0.83 (p=0.006) for the Group-2 cluster.

724 **Panel B. Placement of proviruses on the phylogenetic tree of Thermotogota genomes**

725 **reconstructed from 74 single copy protein-coding genes.** Closely related genomes

726 (distance > 0.1), monophyletic genomes from the same genus, and clades consisting of

727 only metagenome assembled genomes were collapsed. Identified proviruses are indicated

728 next to their respective host genera. The tree was visualized in iTOL [56]. Tree scale,

729 substitutions per site.

730

731 **Supplementary Fig. S2. Placement of complete Thermotogota and Firmicutes**

732 **proviruses on the viral proteomic tree.** The viral proteomic tree is from ViPTree v. 1.9

733 [35], and only the relevant region of the tree is shown. The Thermotogota and Firmicutes

734 proviruses are labeled with red stars. Taxonomy of the related viruses and their hosts is

735 indicated as color bars next to a terminal leaf on the tree.

736

737 **Supplementary Fig. S3. Maximum likelihood trees of three families of terminase**

738 **large subunit genes.** The phylogenetic trees displayed were constructed using RAxML
739 as implemented in Geneious v. 10 with a GAMMA-WAG substitution model and 100
740 bootstrap replicates. The trees should be considered unrooted. Bootstrap support > 70% is
741 shown on branches as circles, with the size corresponding to the strength of support.
742 Taxonomic labels of Thermotogota with proviruses are shown in red bold font, with the
743 provirus name listed after the host name. Thermotogota homologs from genomes with no
744 detected provirus are listed in bold font. Numbers in front of each taxon name represent
745 database accession numbers. The tree was visualized in iTOL and rooted by midpoint
746 rooting and should be considered unrooted [56].

747

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

752

753 **Supplementary Fig. S5. Comparison of the three complete and two incomplete**
754 **Thermotogota Group 2 provirus sequences.** Virus name and the genus the host belongs
755 to is indicated. The regions with significant pairwise BLASTX similarity scores are
756 connected, red indicates that sequence is in the same direction while blue indicates that
757 the similar sequences are on opposite strands. The predicted ORFs are color-coded based
758 on their function and should be considered approximate, because it relies only on gene
759 annotations. Selected gene annotations are included and abbreviated; HMT ATPase:

760 heavy metal translocating ATPase, FMN reductase: flavine mono nucleotide reductase,
761 HAD family phosphatase: haloacid dehalogenase superfamily of hydrolase),
762 dimethyladenosine trf: dimethyladenosine transferase, 2Fe-2S bp: 2Fe-2S binding p
763 rotein, MFS transporter: multi facilitator superfamily transporter. The figure was
764 produced using genoPlotR [27].

765

766 **Supplementary Fig. S6. Maximum likelihood trees of three potential virus receptor**
767 **genes. Panel A: Competence protein ComEA, Panel B: oxaloacetate decarboxylase**
768 **and Panel C: ATP-binding cassette, subfamily B.** Bootstrap support > 70% is shown
769 on branches as circles, with the size corresponding to the strength of support. The names
770 of Thermotogota taxa that contain Group 2 proviruses are displayed in red font and
771 Firmicutes with Group 2-like proviruses are displayed in blue font. Clades containing
772 sequences from the same genus are collapsed into wedges. The trees were rooted using
773 midpoint rooting, and should be considered unrooted. The trees were visualized in iTOL
774 [56].

775

776 **Supplementary Fig. S7. Overview of the alignment of the ABC transporter ATP-**
777 **binding protein in Thermotogota and Firmicutes genomes with Group 2 proviruses.**
778 Sites 100% conserved in all sequences sites are highlighted in color, while variable sites
779 are shown in grey. Transmembrane regions, predicted using the TMHMM Server v. 2.0,
780 are shown in red above the alignment (<http://www.cbs.dtu.dk/services/TMHMM-2.0/>).

781

782 **Supplementary Fig. S8. Maximum likelihood tree of the moron gene annotated as a**
783 **cadmium transporter.** Bootstrap support > 70% is shown on branches as circles, with
784 the size corresponding to the strength of support. Taxon names of Thermotogota with a
785 provirus are given in red and taxon name of Firmicutes with a Group 2-like provirus are
786 given in blue. Provirus name is also indicated. Thermotogota homologs from genomes
787 with no detected provirus, or where the homolog is found outside the provirus region, are
788 given in bold font. Database accession numbers are shown in front of taxonomic names.
789 The tree was rooted by midpoint rooting and visualized in iTOL [56].
790









