Complete Genome Sequence of the Model Halovirus PhiH1 (ΦH1)

Mike Dyall-Smith 1,2, Felicitas Pfeifer 3, Angela Witte 4, Dieter Oesterhelt 1 and Friedhelm Pfeiffer 1,*

1 Computational Biology Group, Max-Planck-Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany; mike.dyallsmith@gmail.com (M.D.-S.); oesterhe@biochem.mpg.de (D.O.)
2 Veterinary Biosciences, Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, VIC 3052, Australia
3 Department of Biology; Microbiology and Archaea, TU Darmstadt, Schnittspahnstrasse 10, 64287 Darmstadt, Germany; pfeifer@bio.tu-darmstadt.de
4 Department of Microbiology, Immunobiology and Genetics, MFPL Laboratories, University of Vienna, Dr. Bohr-Gasse 9, Vienna 1030, Austria; angela.witte@univie.ac.at

* Correspondence: fpf@biochem.mpg.de; Tel.: +49-89-8578-2323

Received: 11 September 2018; Accepted: 8 October 2018; Published: 12 October 2018

Abstract: The halophilic myovirus Halobacterium virus phiH (ΦH) was first described in 1982 and was isolated from a spontaneously lysed culture of Halobacterium salinarum strain R1. Until 1994, it was used extensively as a model to study the molecular genetics of haloarchaea, but only parts of the viral genome were sequenced during this period. Using Sanger sequencing combined with high-coverage Illumina sequencing, the full genome sequence of the major variant (phiH1) of this halovirus has been determined. The dsDNA genome is 58,072 bp in length and carries 97 protein-coding genes. We have integrated this information with the previously described transcription mapping data. PhiH could be classified into Myoviridae Type1, Cluster 4 based on capsid assembly and structural proteins (VIRFAM). The closest relative was Natrialba virus phiCh1 (ϕCh1), which shared 63% nucleotide identity and displayed a high level of gene synteny. This close relationship was supported by phylogenetic tree reconstructions. The complete sequence of this historically important virus will allow its inclusion in studies of comparative genomics and virus diversity.

Keywords: halovirus; virus; halophage; Halobacterium salinarum; Archaea; haloarchaea; halobacteria; genome inversion

1. Introduction

The temperate myovirus Halobacterium virus phiH (ΦH) infects the extremely halophilic archaeon Halobacterium salinarum strain R1 (DSM 671) and was isolated after the spontaneous lysis of a culture of its host [1]. Purified virions require 3.5 M NaCl for stability, have an isometric head of 64 nm diameter and a long, contractile tail (170 × 18 nm) with short tail fibres [1,2]. Virus preparations contain 3 major and 10 minor proteins [3]. The virus genome is linear dsDNA with a G+C content of 64%, contains a pac site, is about 3% terminally redundant and partially circularly permuted, and estimated to be 59 kb in length [4,5]. In the provirus state, the genome is extrachromosomal, covalently closed and circular, and 57 kb in length [4]. While always classified within the Myoviridae, the genus name has changed over the years from phiH-like viruses to PhiHlikevirus, and most recently to Myohalovirus [6,7]. The species name itself has changed from Halobacterium phage phiH to Halobacterium virus phiH [6,7] but for convenience we will refer to it from here onwards simply as phiH, and the analysed variant as phiH1 or halovirus phiH1.
The original lysate of phiH was found to consist of a mixture of several distinct variants that appeared to have arisen from the activity of insertion sequences. The predominant variant, phiH1, was plaque-purified and a restriction map determined [5]. This was used for further study [3]. PhiH1 became a key model in the study of gene expression and regulation in haloarchaea and was instrumental in the development of genetic tools and methods in these extremophiles. Examples include the polyethylene glycol (PEG)-mediated transfection method [8], the pUBP1 cloning/expression vector [9], the identification of archaeal promoters, mapping transcription start and stop sites [10] and the analysis of gene regulation via repression [11,12]. The presence and function of antisense RNA in haloarchaea was first described in this virus [13]. An 11 kb invertible segment of the virus genome, called the L-region, was found to be flanked on one or both sides by the insertion sequence ISH1.8, and could also circularize to form a 12 kb plasmid (including one copy of ISH1.8), with subsequent loss of the remaining phage DNA [14]. A strain carrying this plasmid was immune to infection [14].

Unfortunately, work on phiH stopped in 1994 [15,16] but a related virus, *Natrialba* virus phiCh1 (ϕCh1), was described a few years later [17] and continues to be studied in the Witte laboratory [18,19]. PhiCh1 infects a haloalkaliphilic archaeon, *Natrialba magadii*, and the genomes of both host and virus are fully sequenced [18,20]. The provirus state of phiCh1 corresponds to plasmid pNMAG03 carried by *Nab. magadii*. A full comparison between phiCh1 and phiH1 was prevented as only parts of the phiH genome were ever determined. This deficit also prevented the inclusion of phiH in broad-ranging studies of virus diversity, taxonomy, and evolution. The aim of this study was to complete the phiH1 genome sequence and provide a thorough annotation. This will not only provide a better understanding of the results from previous studies on this virus but also allow complete genomic comparisons with a wealth of other datasets, including other sequenced viruses, haloarchaeal proviruses, metaviromic/metagenomic and environmental RNA sequences.

2. Materials and Methods

2.1. Virus DNA and Sequencing Methods

Purified phiH1 DNA [1] was originally provided to F. Pfeifer by Hans-Peter Klenk while both were working in the department of W. Zillig [21]. The DNA was stored frozen at −80 °C until use. Sequencing was performed in two stages. For the first stage, all available sequences of phiH1 were downloaded from National Center for Biotechnology Information (NCBI) [22] and imported into the Phred–Phrap–Consed package [23]. Overlapping sequences were assembled and primers designed to gather additional sequences using Sanger technology. This consisted either of primer-walking directly on virus DNA, or on polymerase chain reaction (PCR) amplimers, or PCR-sequencing across gaps. The resulting sequence reads were progressively assembled into contigs, base calls inspected manually and corrected where needed, and new primers designed for further rounds of sequencing until all gaps were closed. Except for overlaps, this approach left most of the previously published sequences unchecked.

In the second stage, short-read Illumina HiSeq sequencing of phiH1 DNA was performed (Max-Planck Genome Centre, Cologne, Germany). This returned 243 Mb of high quality sequence data (coverage = 4200-fold). De-novo assembly did not produce a single contig, due to short read-lengths and the presence of repeat sequences within the viral genome, but reads could be confidently mapped to the genome sequence obtained in the first stage (*Map to Reference* option; Geneious mapper method) in order to improve the sequence reliability.

2.2. CRISPR Spacer Searches

The crass v0.3.12 software [24] was used to extract CRISPR spacer sequences from genomic/metagenomic data available at the NCBI SRA database (accessed 27 July 2018) [25], as described previously [26]. These included all available genomes of members of the class Halobacteria,
and metagenomes of hypersaline environments. CRISPR direct repeats (DR) identified by crass were used to search the CRISPRfinder database (accessed 25 July 2018) [27] for haloarchaea with matching or closely matching DR.

2.3. Bioinformatic Methods

Gene annotation used a combination of gene prediction with GeneMarkS-2 [28] and manual refinement using database searches (BLASTp/BLASTn; nr databases) at the NCBI webserver [29]. Repeats were identified by BLASTn, dot-plot comparison in Yass [30], and with tools within the Geneious software suite [31]. Circos plots were performed via the circoletto webserver [32]. Plots are coloured by the ‘score/max’ ratio of tBLASTx bitscores (real score/maximal score). Colours are: blue ≤ 0.25, green ≤ 0.50, orange ≤ 0.75, red > 0.75. Sequence mapping, alignments, editing and phylogenetic tree reconstructions were performed with Geneious software version 10.2 [31]. For phylogenetic tree reconstructions, protein sequences were first aligned using CLUSTALW, and trees inferred using the Neighbor-Joining algorithm (within Geneious). Consensus trees were determined after 100 bootstrap repetitions. Protein structural modelling used the I-Tasser webserver [33]. Identification of the pac site utilised the program PhageTerm [34] as implemented on the CPT Phage Galaxy [35]. The VIRFAM webserver [36] uses proteins of the phage head-neck-tail module to cluster phages into related groups, and was used to classify phiH1.

2.4. Data Availability

The phiH1 genome sequence has been deposited at Genbank under the accession MK002701. Raw reads were submitted to the SRA archive under accession SRP159490.

3. Results and Discussion

3.1. Sequence and Annotation of PhiH1

The previously sequenced regions of the phiH1 genome represented about 50% of the complete sequence (Figure 1, red lines). Using virus DNA as template, the gaps between these sequences were PCR amplified and Sanger sequenced. However, the quality of the previously sequenced regions was of uncertain reliability. High-coverage Illumina sequencing (ca. 4200-fold) was then used to enhance sequence confidence. Sequence revisions were only found to be required in previously deposited sequences but not to Sanger sequencing results of the first stage of the project. While the virus DNA found in capsid particles is linear, the head-full packaging process produces a population of molecules that are terminally redundant and partially circularly permuted [1]. The complete genome sequence determined in the current study is represented as the provirus form; a circular sequence of 58,072 bp. This value is close to the published size of 57 kb, estimated from restriction fragment sizes [4,37]. The G+C content of the genome was 63.7%, almost identical to the published value of 64% [3] but slightly lower than that of the host chromosome (68.0%) [38].

Figure 1. Diagram of the phiH1 genome with lines below showing regions previously sequenced (red) along with their database accessions. The blue lines (NEW) indicate regions sequenced in the present study by Sanger sequencing. Tick marks (dark green) below the blue lines show the positions of oligonucleotide primers used for PCR and primer-walking. Dots at the right and left contig ends indicate sequence continuity between them. Scale bar at top shows position in bp.
The original restriction map of phiH1 DNA, as determined by [5], corresponded closely with the in silico map inferred from the phiH1 genome sequence (Figure S1). The pac site located at the left end of the restriction map matched closely to the corresponding pac sequence of phiCh1. While the pac site of phiCh1 had been localized by restriction mapping [18], it had not been precisely mapped. For consistency, the start point of phiH1 was set to the corresponding start of phiCh1 even though this splits the terS gene. Using this numbering, the program PhageTerm [34] was used to analyse the mapping of Illumina reads to the phiH1 genome, and this located the pac site terminal base at nt 46, with high probability \((p = 2.5 \times 10^{-238})\). This is within the terS coding sequence (CDS) close to the stop codon and within a GC-rich region that is strongly conserved between phiH1 and phiCh1.

Annotation of the phiH1 genome resulted in 97 CDS (Table 1), most of which were encoded on the plus strand (86/97, Figure 2 panel b), and were frequently closely spaced, with 45 overlapping at start/stop codons and 23 separated by \(\leq 8\) nt. Many genes were in functional groupings typical of bacteriophages (Figure 2 panel b). The left end of the genome encodes DNA packaging proteins (e.g., terminase, portal protein), then virus assembly and structural proteins (e.g., major capsid protein, tape-measure protein, tail proteins). The three main proteins of purified virus were originally labelled by their estimated sizes on sodium dodecylsulfate (SDS)-polyacrylamide gels (22, 53 and 80 kDa) [1], which were later revised to 27, 46 and 80 kDa [3] but in 1994, Stolt et al. [39] determined their N-terminal amino acid sequences and used this information to map the proteins (HP20, HP32 and HP67) to their genes (hp20, hp32, hp67) and sequence them. The inferred molecular weights (MWs) of proteins HP20 and HP67 were noted by these authors to be much smaller than previous estimates. In the present study, the locations of these genes on the full genome sequence have been resolved, an error in the hp20 (accession X80161) coding sequence was corrected, and the MWs of the inferred proteins calculated (11.6, 35.4 and 45.5 kDa). For consistency we have retained the original gene names (Table 1).

The next genomic region is a replication/regulatory module (the L-region) that encodes RepR (repressor), a ParA-family protein (partition) and RepH (replication). There is also a VapC-like protein that together with the small overlapping upstream CDS may form a toxin-antitoxin pair that could be involved in plasmid maintenance [40]. The right end of the genome carries many genes with unknown function but includes genes specifying DNA methylases and cell lysis proteins. The taxonomic position of phiH1 was assessed using the VIRFAM webserver [36], which classifies bacteriophages and archaeal viruses based on the order and similarity of capsid assembly/structural proteins. Consistent with previous studies [6], phiH1 was classified by this system as a member of the Myoviridae (Type1, Cluster 4).
Table 1. Annotated coding sequences (CDS) of halovirus phiH1.

<table>
<thead>
<tr>
<th>Start (nt)</th>
<th>Stop (nt)</th>
<th>Locus_tag</th>
<th>Length (bp)</th>
<th>Direction</th>
<th>Gene</th>
<th>Product</th>
<th>Homologs (^1)</th>
<th>Homologs (^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>115</td>
<td>717</td>
<td>PhiH1_005</td>
<td>603</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p02, ORF1</td>
<td>Nmag_4251</td>
</tr>
<tr>
<td>710</td>
<td>2371</td>
<td>PhiH1_010</td>
<td>1662</td>
<td>+</td>
<td>terL</td>
<td>terminase large subunit TerL</td>
<td>PhiCh1p03, ORF2</td>
<td>Nmag_4252</td>
</tr>
<tr>
<td>2377</td>
<td>2505</td>
<td>PhiH1_015</td>
<td>129</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>Nmag_4253</td>
<td></td>
</tr>
<tr>
<td>2498</td>
<td>2689</td>
<td>PhiH1_020</td>
<td>192</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p05, ORF4</td>
<td>Nmag_4255</td>
</tr>
<tr>
<td>2686</td>
<td>4242</td>
<td>PhiH1_025</td>
<td>1557</td>
<td>+</td>
<td>por</td>
<td>portal protein Por</td>
<td>PhiCh1p07, ORF6</td>
<td>Nmag_4257</td>
</tr>
<tr>
<td>4246</td>
<td>5187</td>
<td>PhiH1_030</td>
<td>942</td>
<td>+</td>
<td>-</td>
<td>head morphogenesis protein</td>
<td>PhiCh1p08, ORF7</td>
<td>Nmag_4258</td>
</tr>
<tr>
<td>5261</td>
<td>5587</td>
<td>PhiH1_035</td>
<td>327</td>
<td>+</td>
<td>hp20</td>
<td>capsid protein HP20</td>
<td>[AJF28118.1]</td>
<td></td>
</tr>
<tr>
<td>5667</td>
<td>7466</td>
<td>PhiH1_040</td>
<td>1800</td>
<td>+</td>
<td>-</td>
<td>prohead protease</td>
<td>(^4) PhiCh1p09, ORF8</td>
<td>(^4) PhiCh1p10, ORF9</td>
</tr>
<tr>
<td>7506</td>
<td>8468</td>
<td>PhiH1_045</td>
<td>963</td>
<td>+</td>
<td>hp32</td>
<td>major capsid protein HP32</td>
<td>PhiCh1p12, ORF11</td>
<td>Nmag_4260</td>
</tr>
<tr>
<td>8481</td>
<td>8933</td>
<td>PhiH1_050</td>
<td>453</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p13, ORF12</td>
<td>Nmag_4261</td>
</tr>
<tr>
<td>8940</td>
<td>9542</td>
<td>PhiH1_055</td>
<td>603</td>
<td>+</td>
<td>ada</td>
<td>head-tail adaptor protein Ada</td>
<td>PhiCh1p14, ORF13</td>
<td>Nmag_4262</td>
</tr>
<tr>
<td>9539</td>
<td>9919</td>
<td>PhiH1_060</td>
<td>381</td>
<td>+</td>
<td>hco</td>
<td>head closure protein type 1 Hco</td>
<td>PhiCh1p15, ORF14</td>
<td>Nmag_4263</td>
</tr>
<tr>
<td>9921</td>
<td>10,202</td>
<td>PhiH1_065</td>
<td>282</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p16, ORF15</td>
<td>Nmag_4264</td>
</tr>
<tr>
<td>10,202</td>
<td>10,636</td>
<td>PhiH1_070</td>
<td>435</td>
<td>+</td>
<td>nep</td>
<td>probable neck protein type 1 Nep</td>
<td>PhiCh1p17, ORF16</td>
<td>Nmag_4265</td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>Start (nt)</th>
<th>Stop (nt)</th>
<th>Locus_tag</th>
<th>Length (bp)</th>
<th>Direction</th>
<th>Gene</th>
<th>Product</th>
<th>Homologs 1: phiCh1, ORF pNMAG03 [Other]</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,643</td>
<td>11,239</td>
<td>PhiH1_075</td>
<td>597</td>
<td>+</td>
<td>tco</td>
<td>tail completion protein type 1 Tco</td>
<td>PhiCh1p18, ORF17 Nmag_4266</td>
</tr>
<tr>
<td>11,259</td>
<td>12,557</td>
<td>PhiH1_080</td>
<td>1299</td>
<td>+</td>
<td>hp67</td>
<td>tail sheath protein HP67</td>
<td>PhiCh1p19, ORF18 Nmag_4267</td>
</tr>
<tr>
<td>12,607</td>
<td>13,002</td>
<td>PhiH1_085</td>
<td>396</td>
<td>+</td>
<td>-</td>
<td>probable structural protein</td>
<td>PhiCh1p20, ORF19 Nmag_4268</td>
</tr>
<tr>
<td>13,006</td>
<td>13,407</td>
<td>PhiH1_090</td>
<td>402</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p21, ORF20 Nmag_4269</td>
</tr>
<tr>
<td>13,572</td>
<td>13,745</td>
<td>PhiH1_095</td>
<td>174</td>
<td>-</td>
<td>-</td>
<td>DUF4177 domain protein</td>
<td>[SEH60446.1]</td>
</tr>
<tr>
<td>13,792</td>
<td>16,581</td>
<td>PhiH1_100</td>
<td>2790</td>
<td>+</td>
<td>tpm</td>
<td>tape-measure tail protein Tpm</td>
<td>4 PhiCh1p23, ORF22 4 PhiCh1p24, ORF23 Nmag_4272</td>
</tr>
<tr>
<td>16,583</td>
<td>17,104</td>
<td>PhiH1_105</td>
<td>522</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p25, ORF24 Nmag_4273</td>
</tr>
<tr>
<td>17,108</td>
<td>17,446</td>
<td>PhiH1_110</td>
<td>339</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p26, ORF25 Nmag_4274</td>
</tr>
<tr>
<td>17,450</td>
<td>18,298</td>
<td>PhiH1_115</td>
<td>849</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p27, ORF26 Nmag_4275</td>
</tr>
<tr>
<td>18,306</td>
<td>18,446</td>
<td>PhiH1_120</td>
<td>141</td>
<td>+</td>
<td>-</td>
<td>CxxC motif protein [SEH61109.1]</td>
<td></td>
</tr>
<tr>
<td>18,443</td>
<td>18,988</td>
<td>PhiH1_125</td>
<td>546</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p29, ORF28 Nmag_4276</td>
</tr>
<tr>
<td>18,988</td>
<td>19,146</td>
<td>PhiH1_130</td>
<td>159</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td></td>
</tr>
<tr>
<td>19,143</td>
<td>19,508</td>
<td>PhiH1_135</td>
<td>366</td>
<td>+</td>
<td>-</td>
<td>virus-related protein [AGM10900.1]</td>
<td></td>
</tr>
<tr>
<td>19,505</td>
<td>19,867</td>
<td>PhiH1_140</td>
<td>363</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p30, ORF29 Nmag_4277</td>
</tr>
<tr>
<td>19,874</td>
<td>21,148</td>
<td>PhiH1_145</td>
<td>1275</td>
<td>+</td>
<td>bj</td>
<td>baseplate J family protein Bj</td>
<td>PhiCh1p31, ORF30 Nmag_4278</td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>Start (nt)</th>
<th>Stop (nt)</th>
<th>Locus_tag</th>
<th>Length (bp)</th>
<th>Direction</th>
<th>Gene</th>
<th>Product</th>
<th>Homologs $^1$: phiCh1, ORF pN MAG03 [Other]</th>
</tr>
</thead>
<tbody>
<tr>
<td>21,135</td>
<td>22,277</td>
<td>PhiH1_150</td>
<td>1143</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p32, ORF31 Nmag_4279</td>
</tr>
<tr>
<td>22,295</td>
<td>22,678</td>
<td>PhiH1_155</td>
<td>384</td>
<td>+</td>
<td>-</td>
<td>virus-related protein</td>
<td>[AFH21897.1]</td>
</tr>
<tr>
<td>22,683</td>
<td>23,249</td>
<td>PhiH1_160</td>
<td>567</td>
<td>+</td>
<td>-</td>
<td>virus-related protein</td>
<td>[AFH21653.1]</td>
</tr>
<tr>
<td>23,252</td>
<td>25,504</td>
<td>PhiH1_165</td>
<td>2253</td>
<td>+</td>
<td>-</td>
<td>repeat-containing tail fibre protein</td>
<td>PhiCh1p37, ORF36 Nmag_4282 PhiCh1p35, ORF34 Nmag_4286</td>
</tr>
<tr>
<td>25,506</td>
<td>25,787</td>
<td>PhiH1_170</td>
<td>282</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>Nmag_4285</td>
</tr>
<tr>
<td>25,825</td>
<td>26,499</td>
<td>PhiH1_175</td>
<td>675</td>
<td>+</td>
<td>int1</td>
<td>tyrosine integrase/recombinase Int1</td>
<td>PhiCh1p36, ORF35 Nmag_4284</td>
</tr>
<tr>
<td>26,490</td>
<td>26,792</td>
<td>PhiH1_180</td>
<td>303</td>
<td>–</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>Nmag_4283</td>
</tr>
<tr>
<td>26,798</td>
<td>27,766</td>
<td>PhiH1_185</td>
<td>969</td>
<td>–</td>
<td>-</td>
<td>repeat-containing tail fibre protein $^2$</td>
<td>PhiCh1p37, ORF36 Nmag_4282 PhiCh1p35, ORF34 Nmag_4286</td>
</tr>
<tr>
<td>27,803</td>
<td>28,150</td>
<td>PhiH1_190</td>
<td>348</td>
<td>+</td>
<td>-</td>
<td>YncB-like endonuclease</td>
<td>[AGM11801.1]</td>
</tr>
<tr>
<td>28,153</td>
<td>28,386</td>
<td>PhiH1_195</td>
<td>234</td>
<td>+</td>
<td>-</td>
<td>virus-related protein</td>
<td>[AGC34510.1]</td>
</tr>
<tr>
<td>28,379</td>
<td>28,675</td>
<td>PhiH1_200</td>
<td>297</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>[EMA49173.1]</td>
</tr>
<tr>
<td>28,682</td>
<td>28,783</td>
<td>PhiH1_205</td>
<td>102</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
</tr>
<tr>
<td>28,788</td>
<td>29,357</td>
<td>PhiH1_210</td>
<td>570</td>
<td>+</td>
<td>-</td>
<td>transmembrane domain protein</td>
<td>-</td>
</tr>
<tr>
<td>29,394</td>
<td>29,642</td>
<td>PhiH1_215</td>
<td>249</td>
<td>–</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
</tr>
<tr>
<td>29,651</td>
<td>29,941</td>
<td>PhiH1_220</td>
<td>291</td>
<td>–</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p40, ORF39 Nmag_4289</td>
</tr>
<tr>
<td>30,104</td>
<td>30,244</td>
<td>PhiH1_225</td>
<td>144</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>Start (nt)</th>
<th>Stop (nt)</th>
<th>Locus_tag</th>
<th>Length (bp)</th>
<th>Direction</th>
<th>Gene</th>
<th>Product</th>
<th>Homologs</th>
<th>Homologs</th>
</tr>
</thead>
<tbody>
<tr>
<td>30,250</td>
<td>30,414</td>
<td>PhiH1_230</td>
<td>165</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p44, ORF43 Nmag_4292</td>
<td></td>
</tr>
<tr>
<td>30,411</td>
<td>30,806</td>
<td>PhiH1_235</td>
<td>396</td>
<td>+</td>
<td>-</td>
<td>VapC family toxin</td>
<td>PhiCh1p45, ORF44 Nmag_4293</td>
<td></td>
</tr>
<tr>
<td>30,803</td>
<td>31,465</td>
<td>PhiH1_240</td>
<td>663</td>
<td>−</td>
<td>int2</td>
<td>tyrosine integrase/recombinase Int2</td>
<td>PhiCh1p46, ORF45 Nmag_4294</td>
<td></td>
</tr>
<tr>
<td>31,680</td>
<td>31,934</td>
<td>PhiH1_245</td>
<td>255</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>31,939</td>
<td>32,271</td>
<td>PhiH1_250</td>
<td>333</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>Nmag_4297</td>
<td></td>
</tr>
<tr>
<td>32,420</td>
<td>32,857</td>
<td>PhiH1_255</td>
<td>438</td>
<td>−</td>
<td>-</td>
<td>HNH-type endonuclease</td>
<td>PhiCh1p48, ORF47 Nmag_4296</td>
<td></td>
</tr>
<tr>
<td>32,854</td>
<td>33,255</td>
<td>PhiH1_260</td>
<td>402</td>
<td>−</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>[ELY96531.1]</td>
<td></td>
</tr>
<tr>
<td>33,248</td>
<td>34,024</td>
<td>PhiH1_265</td>
<td>777</td>
<td>−</td>
<td>-</td>
<td>parA domain protein</td>
<td>PhiCh1p47, ORF46 Nmag_4295</td>
<td></td>
</tr>
<tr>
<td>34,161</td>
<td>34,430</td>
<td>PhiH1_270</td>
<td>270</td>
<td>−</td>
<td>repR</td>
<td>repressor protein RepR</td>
<td>5 PhiCh1p49, ORF48 Nmag_4298 [ELZ06324.1]</td>
<td></td>
</tr>
<tr>
<td>34,730</td>
<td>35,071</td>
<td>PhiH1_275</td>
<td>342</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>35,068</td>
<td>35,424</td>
<td>PhiH1_280</td>
<td>357</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p50, ORF49</td>
<td></td>
</tr>
<tr>
<td>35,381</td>
<td>38,167</td>
<td>PhiH1_285</td>
<td>2787</td>
<td>+</td>
<td>repH</td>
<td>plasmid replication protein RepH</td>
<td>4 PhiCh1p54, ORF53 4 PhiCh1p55, ORF54 Nmag_4299</td>
<td></td>
</tr>
<tr>
<td>38,262</td>
<td>38,489</td>
<td>PhiH1_290</td>
<td>228</td>
<td>−</td>
<td>imm</td>
<td>probable immunity protein Imm</td>
<td>PhiCh1p56, ORF55 Nmag_4300</td>
<td></td>
</tr>
<tr>
<td>38,733</td>
<td>39,263</td>
<td>PhiH1_295</td>
<td>531</td>
<td>+</td>
<td>-</td>
<td>transcriptional regulator, PadR-like family</td>
<td>PhiCh1p57, ORF56 Nmag_4301</td>
<td></td>
</tr>
<tr>
<td>39,260</td>
<td>39,385</td>
<td>PhiH1_300</td>
<td>126</td>
<td>+</td>
<td>-</td>
<td>CxxC motif protein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>39,382</td>
<td>39,978</td>
<td>PhiH1_305</td>
<td>597</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p59, ORF58 Nmag_4303</td>
<td></td>
</tr>
<tr>
<td>Start (nt)</td>
<td>Stop (nt)</td>
<td>Locus_tag</td>
<td>Length (bp)</td>
<td>Direction</td>
<td>Gene</td>
<td>Product</td>
<td>Homologs $^1$: phiCh1, ORF pNMAG03 [Other]</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>-----------</td>
<td>-------------</td>
<td>-----------</td>
<td>------</td>
<td>---------</td>
<td>------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>39,975</td>
<td>40,133</td>
<td>PhiH1_310</td>
<td>159</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>40,153</td>
<td>40,902</td>
<td>PhiH1_315</td>
<td>750</td>
<td>+</td>
<td>$pcnA$</td>
<td>DNA polymerase sliding clamp PcnA</td>
<td>PhiCh1p60, ORF59 Nmag_4211</td>
<td></td>
</tr>
<tr>
<td>40,908</td>
<td>41,339</td>
<td>PhiH1_320</td>
<td>432</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p61, ORF60 Nmag_4212</td>
<td></td>
</tr>
<tr>
<td>41,339</td>
<td>41,554</td>
<td>PhiH1_325</td>
<td>216</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p62, ORF61 Nmag_4213</td>
<td></td>
</tr>
<tr>
<td>41,547</td>
<td>42,041</td>
<td>PhiH1_330</td>
<td>495</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>42,098</td>
<td>42,490</td>
<td>PhiH1_335</td>
<td>393</td>
<td>+</td>
<td>$tnpA$</td>
<td>IS200-type transposase TnpA</td>
<td>[CAP12925.1]</td>
<td></td>
</tr>
<tr>
<td>42,492</td>
<td>43,748</td>
<td>PhiH1_340</td>
<td>1257</td>
<td>+</td>
<td>$tnpB$</td>
<td>IS1341-type transposase TnpB</td>
<td>[CAP12926.1]</td>
<td></td>
</tr>
<tr>
<td>43,808</td>
<td>44,014</td>
<td>PhiH1_345</td>
<td>207</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>44,007</td>
<td>44,234</td>
<td>PhiH1_350</td>
<td>228</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p66, ORF65 Nmag_4217</td>
<td></td>
</tr>
<tr>
<td>44,231</td>
<td>44,656</td>
<td>PhiH1_355</td>
<td>426</td>
<td>+</td>
<td>-</td>
<td>CxxC motif protein</td>
<td>PhiCh1p68, ORF67 Nmag_4219</td>
<td></td>
</tr>
<tr>
<td>44,646</td>
<td>45,026</td>
<td>PhiH1_360</td>
<td>381</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p69, ORF68 Nmag_4220</td>
<td></td>
</tr>
<tr>
<td>45,023</td>
<td>45,646</td>
<td>PhiH1_365</td>
<td>624</td>
<td>+</td>
<td>-</td>
<td>HNH-type endonuclease</td>
<td>[KYG11427.1]</td>
<td></td>
</tr>
<tr>
<td>45,639</td>
<td>45,926</td>
<td>PhiH1_370</td>
<td>288</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p71, ORF70 Nmag_4222</td>
<td></td>
</tr>
<tr>
<td>45,919</td>
<td>46,350</td>
<td>PhiH1_375</td>
<td>432</td>
<td>+</td>
<td>-</td>
<td>DUF4326 domain protein</td>
<td>PhiCh1p72, ORF71 Nmag_4223</td>
<td></td>
</tr>
<tr>
<td>46,343</td>
<td>46,441</td>
<td>PhiH1_380</td>
<td>99</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>46,438</td>
<td>46,884</td>
<td>PhiH1_385</td>
<td>447</td>
<td>+</td>
<td>-</td>
<td>CxxC motif protein</td>
<td>PhiCh1p74, ORF73 Nmag_4225</td>
<td></td>
</tr>
<tr>
<td>46,865</td>
<td>47,038</td>
<td>PhiH1_390</td>
<td>174</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>$^5$ PhiCh1p73, ORF72 $^5$ Nmag_4224</td>
<td></td>
</tr>
</tbody>
</table>
Table 1. Cont.

<table>
<thead>
<tr>
<th>Start (nt)</th>
<th>Stop (nt)</th>
<th>Locus_tag</th>
<th>Length (bp)</th>
<th>Direction</th>
<th>Gene</th>
<th>Product</th>
<th>Homologs 1: phiCh1, ORF pNMAG03 [Other]</th>
</tr>
</thead>
<tbody>
<tr>
<td>47,031</td>
<td>47,447</td>
<td>PhiH1_395</td>
<td>417</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
</tr>
<tr>
<td>47,440</td>
<td>47,739</td>
<td>PhiH1_400</td>
<td>300</td>
<td>+</td>
<td>-</td>
<td>NTPase protein</td>
<td>[PLX87675.1]</td>
</tr>
<tr>
<td>47,732</td>
<td>49,618</td>
<td>PhiH1_405</td>
<td>1887</td>
<td>+</td>
<td>dcm5</td>
<td>C-5 cytosine-specific DNA methylase Dcm5</td>
<td>$^5$ PhiCh1p81, ORF80 [PCR88664.1]</td>
</tr>
<tr>
<td>49,611</td>
<td>49,931</td>
<td>PhiH1_410</td>
<td>321</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p82, ORF81 Nmag_4234</td>
</tr>
<tr>
<td>49,918</td>
<td>50,037</td>
<td>PhiH1_415</td>
<td>120</td>
<td>+</td>
<td>-</td>
<td>CxxC motif protein</td>
<td>-</td>
</tr>
<tr>
<td>50,091</td>
<td>51,452</td>
<td>PhiH1_420</td>
<td>1362</td>
<td>+</td>
<td>yhdJ</td>
<td>DNA methylase N-4/N-6 domain protein YhdJ</td>
<td>PhiCh1p83, ORF82 Nmag_4235</td>
</tr>
<tr>
<td>51,449</td>
<td>52,024</td>
<td>PhiH1_425</td>
<td>576</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p84, ORF83 Nmag_4236</td>
</tr>
<tr>
<td>52,021</td>
<td>52,791</td>
<td>PhiH1_430</td>
<td>771</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p85, ORF84 Nmag_4237</td>
</tr>
<tr>
<td>52,784</td>
<td>53,152</td>
<td>PhiH1_435</td>
<td>369</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p88, ORF87 Nmag_4240</td>
</tr>
<tr>
<td>53,145</td>
<td>53,504</td>
<td>PhiH1_440</td>
<td>360</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p89, ORF88 Nmag_4241</td>
</tr>
<tr>
<td>53,788</td>
<td>54,369</td>
<td>PhiH1_445</td>
<td>582</td>
<td>+</td>
<td>-</td>
<td>CxxC motif protein</td>
<td>PhiCh1p90, ORF89 Nmag_4242</td>
</tr>
<tr>
<td>54,403</td>
<td>54,771</td>
<td>PhiH1_450</td>
<td>369</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>PhiCh1p91, ORF90 Nmag_4243</td>
</tr>
<tr>
<td>54,794</td>
<td>55,147</td>
<td>PhiH1_455</td>
<td>354</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td>-</td>
</tr>
<tr>
<td>55,144</td>
<td>55,401</td>
<td>PhiH1_460</td>
<td>258</td>
<td>+</td>
<td>-</td>
<td>transmembrane domain protein</td>
<td>PhiCh1p93, ORF92 Nmag_4244</td>
</tr>
<tr>
<td>55,394</td>
<td>55,729</td>
<td>PhiH1_465</td>
<td>336</td>
<td>+</td>
<td>-</td>
<td>transmembrane domain protein</td>
<td>$^3$ PhiCh1p94, ORF93 Nmag_4245</td>
</tr>
<tr>
<td>55,794</td>
<td>57,053</td>
<td>PhiH1_470</td>
<td>1260</td>
<td>+</td>
<td>ycdA</td>
<td>DNA methylase N-4/N-6 domain protein YcdA</td>
<td>PhiCh1p95, ORF94 Nmag_4246</td>
</tr>
<tr>
<td>Start (nt)</td>
<td>Stop (nt)</td>
<td>Locus_tag</td>
<td>Length (bp)</td>
<td>Direction</td>
<td>Gene</td>
<td>Product</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>-----------</td>
<td>-------------</td>
<td>-----------</td>
<td>------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>57,046</td>
<td>57,564</td>
<td>PhiH1_475</td>
<td>519</td>
<td>+</td>
<td>-</td>
<td>uncharacterized protein</td>
<td></td>
</tr>
<tr>
<td>57,621</td>
<td>57,830</td>
<td>PhiH1_480</td>
<td>210</td>
<td>+</td>
<td>-</td>
<td>CxxC motif protein</td>
<td></td>
</tr>
<tr>
<td>57,827&gt;</td>
<td>&lt;63</td>
<td>PhiH1_485</td>
<td>309</td>
<td>+</td>
<td>terS</td>
<td>terminase small subunit TerS</td>
<td></td>
</tr>
</tbody>
</table>

**Homologs**

1. PhiCh1/pNMAG03 homologs of phiH1 proteins show BLASTp E-values < 10^{-20}. For phiCh1 proteins, both the PhiCh1p and originally assigned ORF codes (ORF for open reading frame) are shown (e.g., PhiCh1p02, ORF1). Codes starting with ORF represent the original annotation of the phiCh1 genome [17] (GB accession AF440695.1); and codes starting with PhiCh1p represent the RefSeq version of the annotation of the same genome sequence (GB accession NC_004084). The number shift is due to the terS gene, the N-terminal part being encoded at the end of the genome, and the C-terminal part at its beginning. This ORF is complete in the provirus state due to circularization and in the linear virus state due to terminal redundancy. This gene is ORF98 in the original annotation and PhiCh1p01 in the RefSeq annotation. Codes starting with Nmag represent the annotation of the Natrialba magadii plasmid pNMAG03 [20] (accession CP001935.1). The point of ring opening in pNMAG03 was set between Nmag_4303 and Nmag_4211. Codes in square brackets represent NCBI accessions referring to homologous proteins (BLASTp E-values ≤ 10^{-11}), which are from other sources.

2. Gene PhiH1_185 is encoded on an invertible segment. In the current sequence version, it is inactivated because it is uncoupled from a start codon. By genome inversion, it becomes activated while its partner gene PhiH1_165 becomes inactivated. Overall, this results in tail fibre protein switching.

3. This protein (PhiH1_465) has three predicted transmembrane domains and has been suspected to function as a holin [18].

4. In these cases, the phiCh1 gene is split into two CDS but is continuous in phiH1. These proteins are more distantly related (show less than 39% sequence identity or fall above BLASTp E-values of 10^{-20}). In these cases, a similar genetic context supports their stated relationship.
A GC-profile plot [41] of the phiH1 genome shows a major low point inflection within the L-region (Figure 2, panel a), indicating a potential replication origin. The L-region is ~12 kb in length, can replicate as a plasmid in *Halobacterium* [14], and carries genes encoding a replication protein (RepH), and a DNA-binding repressor (RepR). It can also provide cells with immunity to infection by phiH1 virus. The transcription program of phiH1 during lytic growth (panels c, d and e) has been summarized from previous studies, and shows temporal changes (early, middle and late transcripts). The broad directions of transcription reflect the closely spaced and similarly directed gene clusters as well as the correspondence with functional gene groupings (panel b). The lowest two panels (d, e) summarize the results of hybridizing labelled transcripts from infected cells to Southern blots of restricted phiH1 DNA [42], so mapping transcripts to fragments of the virus genome. Panel c shows a summary of the virus-specific transcripts that were sized by agarose gel electrophoresis and had 5′ start sites mapped. While transcription across the L-region has been examined in more detail compared to the rest of the genome, there remains much that is incomplete or uncertain. For example, the 3′ end of the late transcript labelled T_{LL}, which is depicted ending in a dotted line and question mark (at ~21 kb), has not been determined. This transcript could potentially extend for another 5.5 kb. Counter-transcripts are commonly produced by prokaryotes and their viruses and play important roles in gene regulation. Their presence and activity in phiH1 gene expression has been studied and was one of the first reports of antisense RNA in Archaea [13]. However, this interesting topic remains to be fully explored.

![Figure 2. PhiH1 GC-profile, genetic map, and corresponding transcription program (adapted from [16,42]).](image-url)

(a) GC-profile of the phiH1 genome. (b) Genetic map of the phiH1 genome, showing coding sequences as red, grey or blue arrows. Dotted lines above indicate gene clusters involved in particular functions. Some CDS are labelled above the map, e.g., TerL, terminase large subunit; Portal, portal protein; Tape, tape-measure protein; RepH, replicase (label within CDS arrow); Mt, DNA methylases; TerS, terminase small subunit. Some genes are shown below the map, such as *hp32*, encoding the major capsid protein HP32. Panels c, d and e summarise transcription data from previously published studies, and above them is a colour key that indicates the time of appearance of early (0–1 h, blue), middle (1–2 h, green) and late (>2 h, pink) transcripts. (c) Precise mapping of viral transcripts, including start and termination sites [16,39]. (d) Summary transcription program of lytic
infection based on hybridisation of labelled infected-cell transcripts to restriction fragments of virus DNA [42]. Thin coloured lines indicate whether continuing transcription persists over time. (e) The transcription map data of [42] are shown, projected onto the in silico restriction map of phiH1, as determined from the complete genome sequence (this study). Enzymes are indicated at the left. Numbers on the restriction map refer to those of the original publication of [42] (see also Figure S1). Coloured shading follows that of panels c and d. Dotted pattern shown beyond the right-hand pac site indicates terminus redundancy of virus DNA. Scale bars (in kb) are shown below panels a and e.

 Corrections to the previously sequenced regions resulted in significant changes to several coding sequences. For example, the *tnpB* gene of transposon ISH1.8 (nt 41,906–43,789) was thought to be inactive as it was split into three CDS by multiple mutations [43]. The high-quality Illumina sequence data show, however, that the gene is intact and that the previously reported transposon ISH1.8 (X00805) is actually an exact copy of transposon ISH12 from the host *Hbt. salinarum* strain R1 [44]. The element plays a key role in the mobilisation of the L-region of the genome to form the 12 kb plasmid, pΦH1 [14,43]. Another case is the Dcm5 cytosine methylase, which was also reported as being split [39]. The revised sequence shows that the gene codes for a single, probably functional protein (PhiH1_405, nt 47,732–49,618) and not for the two parts (dcm5a, dcm5b) as previously reported. Although phiH1 carries three potentially active DNA methylase genes (*dcm5*, *yhdJ* and *ycdA*), the presence of modified bases in phiH1 DNA was not detected in the chromatographic (high-pressure liquid chromatography) profiles of deoxyribonucleosides released by enzymatic hydrolysis [45]. In that study, the genomes of phiH and another, unrelated virus (phiN) were analysed, and while unmodified dC was detected in phiH, the phiN genome contained only methylated dC (Figure 3 in [45]). The related halovirus phiCh1 carries homologs of two of the phiH1 methylases, and one of them, N6-adenine methylase (ORF94/M-ϕCh1-I, corresponding to YcdA of phiH1) has been shown to methylate DNA at GATC motifs [46] but the proportion of sites found to be modified in virus DNA by M-ϕCh1-I varies from 5% to 50%, depending upon the infection conditions. Modifying only some of the available sites is presumably advantageous to avoid host restriction, as distinct enzymes may target either unmethylated or methylated sites.

### 3.2. Matches to CRISPR Spacers

The phiH1 genome was used to search for matching CRISPR spacers among metagenomic datasets of hypersaline environments downloaded from the NCBI sequence read archive (SRA; see methods). Only four spacers showing close to moderate similarity to phiH1 were detected (Table 2). These spacers match to virus genes encoding structural and non-structural proteins, and the DRs of these spacers show that they are carried by haloarchaea. The datasets include metagenomes from the USA and Iran, as well as an isolate from the Andaman Islands, India. The results suggest that phiH1-like viruses are geographically widespread.
These spacers match to virus genes encoding structural and non-structural proteins, and the DRs of methods). Only four spacers showing close to moderate similarity to phiH1 were detected (Table 2).

phiH1-like viruses are geographically widespread. USA and Iran, as well as an isolate from the Andaman Islands, India. The results suggest that these spacers show that they are carried by haloarchaea. The datasets include metagenomes from the hypersaline environments downloaded from the NCBI sequence read archive (SRA; see 3).

### Table 2. CRISPR spacers matching phiH1.

<table>
<thead>
<tr>
<th>No.</th>
<th>CRISPR Spacer Matches to phiH1</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DR: GCTTCAACCCCACGGGTTCGCTCTAAAC (Haloarcula/Halorhabdus/Natronomonas)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>phiH1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRJNA337743</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18677 CACGCCAGAGCCGCTCGAGAGAGATCCACG 18712</td>
<td></td>
</tr>
<tr>
<td></td>
<td>phiH1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRJEB18068</td>
<td></td>
</tr>
<tr>
<td></td>
<td>25264 CTCACGAAGGCGAGCAGCCTCCAGCGTCTTAGA 25296</td>
<td></td>
</tr>
<tr>
<td></td>
<td>phiH1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRJEB18068</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18677 CACGCCAGAGCGAGCCTCCAGCGTCTTAGA 18712</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>phiH1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRJEB18068</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18677 CACGCCAGAGCGAGCCTCCAGCGTCTTAGA 18712</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>phiH1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>JMIPO2000025.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>40543 GTCTCGAGGGCGGCACTCGACCGCGATCGACGCG 40580</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>phiH1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>PRJNA337743</td>
<td></td>
</tr>
<tr>
<td></td>
<td>50186 CATGTGTGATGACCTGCCTCCGTAATCGGTCG 50220</td>
<td></td>
</tr>
</tbody>
</table>

The matching spacer sequences were found in the following NCBI bioprojects using the crass program: PRJNA337743, (SRA SRR4030040); Alviso Ponds, San Francisco, CA, USA; metagenome); PRJNA245787 (Halostagnicola sp. A56 26 genome; Andaman Islands, India); PRJEB18068 (Lake Meyghan, Iran; metagenome). Aligned sequences show nt positions for phiH1, and asterisks indicated identical bases. DR: direct repeat (with haloarchaea containing most closely matching DR shown in brackets). Symbols under alignment (\*:\*) indicate identical, similar and weakly similar residues, respectively (based on Gonnet PAM 250 matrix).
3.3. Relatives and Phylogeny of PhiH1

The only close matches to the phiH1 genome in the GenBank database were phiCh1 and the corresponding *Nab. magadii* plasmid pNMAG03 (BLASTn, accessed 20 July 2018). A dot-plot comparison of phiH with phiCh1 (Figure S2) revealed a largely colinear relationship (green line) and an overall nucleotide similarity of 63%. The plot also highlights several indels (line gaps) and two regions showing inversions (red lines). Inversion 1 (nt 24,227–27,767) corresponds closely in sequence and arrangement to the invertible region described in phiCh1 (ORF34-36) that has a central XerD type integrase/recombinase gene flanked by inverted repeats, and facilitates switching between two related tail fibre genes, each containing numerous short repeats [18]. The phiH1 orthologous integrase is PhiH1_175. In the current sequence version, PhiH1_165 is active while PhiH1_185 is uncoupled from a start codon and thus is inactivated. Upon inversion of the genome segment, PhiH1_185 is activated while PhiH1_165 is inactivated. Overall, this results in tail fibre protein switching, which may affect receptor binding specificity and host range of phiH1. The similarity in the tail fibre protein repeats is high enough to be detectable at the DNA level, which results in the X-shaped pattern for this region in the dot-plot. Inversion 2 (nt 31,932–34,126) occurs within the phiH1 L-segment, encompasses four CDS including a ParA-domain protein, and is nearby a different integrase/recombinase gene (Int2, PhiH1_240). Protein searches (BLASTp) of the phiH1 genome returned matches to phiCh1, a limited number of haloarchaeal genomes (often 5–10, which may flag proviral regions) and the haloarchaeal caudoviruses BJ1 [47] and CGphi46 (NC_021537), both of which infect *Halorubrum* spp. Pairwise alignments (BLASTp) between all phiH1 and matching phiCh1 proteins gave an average protein sequence identity of 70% (range 39–95%; with a few exceptions, see footnote 5, Table 1). Figure 3 is a graphical comparison of phiH1 proteins (tBLASTx) with those of phiCh1, BJ1, CGphi46 and, as an outlier, HSTV-1. The *Haloarcula* caudovirus HSTV-1 [48] shows very low similarity to phiH1. The figure summarizes the close similarity of phiH1 and phiCh1 proteins. BJ1 and CGphi46 show far fewer matching regions, mainly to proteins encoded near the left end of the phiH1 genome, a region specifying portal and capsid proteins. The three significant matches to HSTV-1 were to a methyltransferase (HSTV1_52), a hypothetical protein (HSTV1_53), and a DNA polymerase sliding clamp protein (HSTV1_40).

While several caudovirus proteins have been used to infer virus phylogenies, the major capsid protein (MCP) is often used because of its functional constraints maintaining a conserved structure [49]. Figure 4 shows a tree reconstruction using an alignment of phiH1 MCP (HP32, 35.4 kDa) and related sequences. Haloarchaeal proteins are seen to branch together (pink shading) and within this cluster the phiH1 and phiCh1 MCPs form a distinct and closely branching clade. These two proteins share 82% amino acid identity. The MCPs of CGphi46 and BJ1 branch at distant locations from each other and from phiH1 MCP. Structures of close homologs of phiH1 HP32 have not yet been determined. However, the major capsid proteins of bacterial caudoviruses and eukaryotic herpesviruses share a common folding structure, the archetype of which is the phage HK97 MCP (gp5) [50]. Consistent with this, modelling of the phiH1 MCP (I-Tasser) returned bacteriophage HK97 gp5 (PDB 2fs3A) as the closest matching structure (Template Modeling (TM)-score = 0.848, Root-Mean-Square Deviation (RMSD) = 1.17). Based on structure prediction and homology modelling, the HK97-fold may also be present in the MCP of phiCh1 [49]. The structure of the MCP of the haloarchaeal podovirus, HSTV-1, has recently been shown to be of the HK97 type [48].
PhiH1 and phiCh1 display a close sequence similarity across most of their genomes yet infect physiologically and biochemically different haloarchaeal hosts. *Hbt. salinarum* is a widely distributed neutrophilic heterotroph with glycolipid-containing membranes, and has often been isolated from spoilage of salted products while *Nab. magadii* is a haloalkaliphile (optimum pH 9.5) that lacks glycolipids [51] and is restricted in its distribution to highly alkaline salt lakes [52]. Looking more widely, the presence of phiH1 MCP homologs in diverse genera of haloarchaea and two haloviruses (Figure 4) indicates that the *Myohalovirus* genus and related viruses are a highly successful group, the reasons for which are worthy of more detailed study, particularly when large-scale cultivation of *Halobacterium* becomes more common [53]. PhiH1 has been well studied in the past, and the completion of its genome sequence now allows it to be included in much of the sequence-based studies used today, including comparative virology, detection of proviruses in archaeal genomes, virus evolution and the microbial ecology of hypersaline environments.

Figure 3. Circos plot of amino acid similarity (tBLASTx) between phiH1 and the haloviruses phiCh1, BJ1, CGphi46 and HSTV-1. The threshold for connecting lines was E-value $\leq 10^{-40}$, with line colours reflecting the ratio of actual tBLASTx bitscore to the maximal score (using ‘score/max’ ratio colouring with blue $\leq 0.25$, green $\leq 0.50$, orange $\leq 0.75$, red $> 0.75$). The outer histogram counts how many times each colour has hit the specific part of the sequence and uses an equivalent colouring scheme. The distance between successive tick marks shown along each virus genome represents 0.1 of the full genome length. Protein names shown along the phiH1 genome indicate the positions of the corresponding genes.
Author Contributions: Conceptualization, visualization, investigation, M.D.-S.; data curation, Fr.P., M.D.-S.; writing—review & editing, A.W., D.O., Fe.P., Fr.P., M.D.-S.; validation, A.W.

Funding: This research was funded by the Max Planck Society, Germany, to Dieter Oesterhelt, emeritus Director of the Department of Membrane Biochemistry, Max Planck Institute, Martinsried, Germany.

Conflicts of Interest: The authors declare no conflict of interest.

References


Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/9/10/493/s1, Figure S1: Original phiH1 restriction map compared to in silico map from genome sequence. Figure S2: Dot plot sequence comparison of phiH1 and phiCh1 genomes.

Author Contributions: Conceptualization, visualization, investigation, M.D.-S.; data curation, Fr.P., M.D.-S.; funding acquisition, D.O., Fr.P.; project administration, Fr.P.; resources, A.W., Fe.P.; writing—original draft, Fr.P., M.D.-S.; writing—review & editing, A.W., D.O., Fe.P., Fr.P., M.D.-S.; validation, A.W.

Funding: This research was funded by the Max Planck Society, Germany, to Dieter Oesterhelt, emeritus Director of the Department of Membrane Biochemistry, Max Planck Institute, Martinsried, Germany.

Conflicts of Interest: The authors declare no conflict of interest.

Figure 4. Phylogenetic tree reconstruction (NJ method) of major capsid proteins (MCP) of phiH1, other haloviruses and related proteins of haloarchaea. Species names of haloarchaeal species are shown, with accession numbers given at the right side. Bootstrap confidence values (100 repetitions) are shown at branch points. The pink shading highlights taxa belonging to the class Halobacteria. Scale bar (expected changes per site) is shown at top. The outgroup (not shown) consisted of distantly related MCP sequences of Bacillus spp. (WP_001060157.1, WP_098773561.1, WP_001064748.1 and WP_000178926.1).
