Genome Annotation and Intraviral Interactome for the *Streptococcus pneumoniae* Virulent Phage Dp-1

Mourad Sabri,† Roman Häuser,2,7† Marc Ouellette,3 Jing Liu,4‡ Mohammed Dehbi,4§ Greg Moeck,4 Ernesto García,5 Björn Titz,6 Peter Uetz,7* and Sylvain Moineau1*

Département de Biochimie, de Microbiologie et Bio-Informatiques, Faculté des Sciences et de Génie, Groupe de Recherche en Écologie Buccale, Faculté de Médecine Dentaire, Félix d’Hérelle Reference Center for Bacterial Viruses, Université Laval, Québec, Canada G1V 0A6; Institute of Toxicology and Genetics, Karlsruhe Institute for Technology, Karlsruhe, Germany; Centre de Recherche en Infectiologie de l’Université Laval, Centre Hospitalier Universitaire de Québec, Québec, Canada G1V 4G2; The Medicines Company, Ville St. Laurent, Quebec, Canada; Centro de Investigaciones Biológicas, CSIC, Madrid, Spain; Crump Institute for Molecular Imaging, Los Angeles, California; and J. Craig Venter Institute, Rockville, Maryland

Received 20 September 2010/Accepted 8 November 2010

*Streptococcus pneumoniae* causes several diseases, including pneumonia, septicemia, and meningitis. Phage Dp-1 is one of the very few isolated virulent *S. pneumoniae* bacteriophages, but only a partial characterization is currently available. Here, we confirmed that Dp-1 belongs to the family *Siphoviridae*. Then, we determined its complete genomic sequence of 56,506 bp. It encodes 72 open reading frames, of which 44 have been assigned a function. We have identified putative promoters, Rho-independent terminators, and several genomic clusters. We provide evidence that Dp-1 may be using a novel DNA replication system as well as redirecting host protein synthesis through queuosine-containing tRNAs. Liquid chromatography-mass spectrometry analysis of purified phage Dp-1 particles identified at least eight structural proteins. Finally, using comprehensive yeast two-hybrid screens, we identified 156 phage protein interactions, and this intraviral interactome was used to propose a structural model of Dp-1.

*Streptococcus pneumoniae* (pneumococcus) is a low-GC-content, Gram-positive bacterium belonging to the mitis group of streptococci (40). It is a facultative human pathogen, colonizing the mucosal surface of the upper respiratory tract and causing invasive infections such as pneumonia, meningitis, and sepsis (97). *S. pneumoniae* diseases are associated with severe morbidity and a high rate of mortality, especially among young children and the elderly. Antibiotics are commonly used to control and cure the invasive pneumococcal diseases, but this high exposure to antibiotics has led to multiresistant *S. pneumoniae* strains worldwide (79). Vaccination is efficient in preventing invasive diseases (103), but pneumococcal clones expressing capsular polysaccharides with serotypes not included in the current vaccines’ formulations exist, and the incidence of invasive diseases induced by these vaccine-escaping clones is increasing (34). Replacement of the vaccine-included clones by the vaccine-escaping ones is possible (91).

The appearance of antibiotic-resistant and vaccine-escaping clones is partly due to the high genome plasticity of *S. pneumoniae*. This genome plasticity is supported by the ability of the pneumococcus to acquire DNA through natural competence and by genome rearrangements facilitated by the mobile elements present in its genome (15, 43, 93). *S. pneumoniae* is also able to colonize many sites inside the human host using phase variation of cell surface components such as polysaccharide capsule, lipoteichoic acid, and cell-surface expressed proteins (42, 78, 101). The complexity of physiology and genetics of *S. pneumoniae* complicates the search for novel therapeutic approaches needed to counter increasing antibiotic resistance and emergence of the vaccine-escaping clones.

One of the novel therapeutic approaches toward antibiotic-resistant strains is the use of bacteriophage proteins to lyse pneumococcal cells. The feasibility of this approach was first demonstrated when bacteriophage endolysin enzymes were used successfully to kill pneumococcal cells (for recent reviews, see references 28, 33, and 54). Further investigation and characterization of pneumococcal bacteriophages might prove useful for the identification of other proteins efficient in therapeutic control of pneumococcus.

The isolation of pneumococcal phages has been previously reported, including virulent (63, 75, 85, 94) and temperate phages (8, 51, 68, 69, 81), the latter being by far the most studied in pneumococci. In fact, most clinical *S. pneumoniae* isolates have been shown to carry prophages (77, 82). While...
the complete genomic sequences of several pneumococcus prophages are available in GenBank, only one virulent phage has been sequenced, namely, phage Cp-1, a member of the Podoviridae family (66).

Dp-1 was the first described pneumococcal virulent phage (63). This double-stranded DNA phage belongs to the family Siphoviridae (long noncontractile tail) and has a lipid-containing capsid, adsorbs to phosphorylcholine residues located at the cell surface (55), and apparently needs the host autolytic system for efficient progeny release (84). The Pal endolysin of Dp-1 (27, 90, 98) has been characterized at the molecular level and successfully used against virulent pneumococcus strains in vitro and in vivo (59, 52, 53, 80). Despite the early characterization of the physicochemical properties of phage Dp-1, considerably less is known about its genomic organization, as only a partial sequence is available in patent databases (73) as well as a limited annotation (26). The vast majority of Dp-1 proteins have not been analyzed.

Since most phage proteins are either involved in phage assembly or some other stage of the infection process, we reasoned that a systematic analysis of protein-protein interactions may shed light on the function of the many uncharacterized Dp-1 proteins. Several protein interactomes have been published for bacteria (71, 76, 95), but only one complete phage interactome is available, namely, for Escherichia coli phage T7 (7). To our knowledge, interactomes have not been used to annotate any phage genome.

Here we present the sequence and annotation of the Dp-1 genome based on homology and proteome data. We have identified structural proteins of Dp-1 and systematically analyzed all putative Dp-1 proteins for protein-protein interactions. We have tested nearly all intraviral protein combinations using MS (7). To our knowledge, interactomes have not been used to annotate any phage genome.

**Materials and Methods**

**Microbial strains and growth conditions.** S. pneumoniae avirulent unencapsulated strain R6, a derivative of virulent serotype 2 strain D39 through strain stored every hour. When the OD 600 reached 0.4, usually after 3 h, it was diluted 10-fold. To our knowledge, interactomes have not been used to annotate any phage genome.

**DNA extraction from purified phage Dp-1 samples.** DNase (1 μg) and RNase (1 μg) were added to 200 μl of the purified phage samples prior to 30 min incubation at 37°C, followed by an addition of 200 μg of proteinase K and 1 h of incubation at 65°C. Ammonium acetate (7.5 M) was added at 135 μl, followed by 350 μl of phenol-chloroform. The resulting solution was mixed by inversion. The mix was then centrifuged for 10 min at 22,000 × g at 4°C, and the aqueous phase was carefully separated from the phenol layer. The aqueous phase was then precipitated with 2 volumes of ethanol and incubated at −20°C for 95% ethanol, and centrifuged for 10 min at 22,000 × g at 4°C. The resulting pellet was washed twice with 75% ethanol, air dried, and solubilized in 50 μl of water. The DNA concentration was measured using a Nanodrop spectrophotometer (Thermo).

**Dp-1 sequencing.** Phage DNA (10 μg) was sheared by sonication, repaired with T4 DNA polymerase and Klenow fragments, size selected (1 to 2 kb), and cloned into the HinII site of plasmid pKSII (Stratagene). Individual clones were sequenced by using BigDye primer or BigDye terminator cycle reaction ABI Prism technology, and contigs were assembled using Sequencer 3.1 (Gene Codes, Ann Arbor, MI) or Phred-Phrap/Consed 12.0 (CodonCode, Dedham, MA) software. Primer walking was used to close gaps in the sequence.

**Analyses of Dp-1 genome.** The nucleotide sequence of Dp-1 was analyzed for the presence of open reading frames (ORFs) at least 30 codons using Genemark.hmm (10), the Easygene 1.2 server (46), and the NCBI ORF Finder. Identified ORFs were manually screened for the presence of a Shine-Dalgarno-like sequence (ribosome binding site [RBS]) located within a maximum of 25 nucleotides upstream of the ORFs. When alternative start codons were identified in ORF Finder, the closest start codon to the RBS was selected as the start of the putative gene. Putative genes were compared to genomic databases using PSI-BLAST (2), and FASTA (72). Protein sequences were also analyzed for conserved domains by using CDD (59), InterProScan (66), and SMART HMM (49). Putative promoters were recognized by searching for −35 box (TTGACA) sequences and −10 (TATAA) box sequences upstream of genes or extended −10 promoters (TATGATAAT) (56). The presence of secondary RNA stem-loop structures was investigated using mFOLD (108) to identify putative terminators. Adenine- and thymine-containing regions were also investigated upstream and downstream of these identified structures (48). The Dp-1 genome was also searched for the presence of RNA-encoding sequences using RNASeek (88).

**Identification of structural proteins.** Trypsin digestion of the purified phage Dp-1 followed by liquid chromatography-tandem mass spectrometry (LC/MS-MS) was performed at the proteomic platform of the Centre de Génomique du Québec (Quebec City, Quebec, Canada). The resulting peptide sequences were searched against deduced gene products of Dp-1 and S. pneumoniae strain R6 protein databases.

**Cloning of baits and preys.** The ORFs of Dp-1 were PCR amplified with the VentRI DNA Polymerase (New England Biolabs) and cloned via Gateway site-specific recombination. PCR products were cloned simultaneously in pDONR207 and pDONR221 via BP reaction (Invitrogen). Sequence positions of ORFs were used to design gene-specific primers by using the default settings of the Express Primer Tool (19). Forward oligonucleotides included a sequence of the attB1 recombination site (5′-AAAAAAGCAGGCTTA-3′) followed by the
native start codon and the gene-specific priming region; reverse oligonucleotides contained the attB2 recombination site region (5'-AGAAAGCTCTGCTGTTTA-3') with TAA as a stop codon followed by the 3'-terminal ORF priming region. By using the forward and reverse adapter oligonucleotides (5'-GGGGACACCAAGTGTTTCTGATTGGA-3', 5'-GGGGACACCTTTGATCAAGAAGGTGATC-3'), the complete attachment sites were fused to the PCR products during a second PCR. Entry clones were confirmed by sequencing using oligonucleotides for pENT221 (M13 forward, 5'-GTAAACACGCGCCGATG-3'; M13 reverse, 5'-AAGCAGTGATGACGATG-3') and for pENTR207 (forward, 5'-TTAACGCTAGCATGGGATCTC-3', reverse primer for pGBK/GADT7g, 5'-CATCAGAGATTTTGAGGCAC-3'). Y2H expression vectors were obtained by Gateway LR reaction. Destination vectors pDEST32 and pDEST22 (Invitrogen) were simultaneously recombined with individual pENTR221 entry vectors and pGBK7g/pGADT7g (Clontech) (96) with pENTR207 plasmids. After plasmid purification, all expression plasmids were checked for correct insert size by a control PCR using the (Clontech) (96) with pENTR207 plasmids. After plasmid purification, all expression plasmids were checked for correct insert size by a control PCR using the forward and reverse adapter oligonucleotides (5'-TTATCCACA-3', reverse oligonucleotide for pGBK/GADT7g, 5'-TTAACGCTAGCATGGGATCTC-3', reverse primer for pGBK/GADT7g, 5'-CTTTCGTTTAAAACCTAAGGTTCAC-3', for pDEST32/22 plasmids, 5'-AGCGCACAACCTTGATGGAAGAC-3').

Yeast two-hybrid screens. pDEST32/pDEST22 and pGBK7g/pGADT7g bait and prey plasmids were individually transformed into haploid AH109 or Y187, respectively (31), by a standard lithium acetate method. After selection on selective medium, preys were arrayed as quadruplicates to ensure experimental reproducibility. Single pDEST32 and pGBK7g (Invitrogen) were simultaneously recombined with individual pENTR221 entry vectors and pGBK7g/pGADT7g (Clontech) (96) with pENTR207 plasmids. After plasmid purification, all expression plasmids were checked for correct insert size by a control PCR using the forward and reverse adapter oligonucleotides: forward primer for baits, 5'-GGAGAACGAGGTG-AGTTAACAAATCTAAGTG-3'; for preys, 5'-GGATGTTTATACCCACTACATG-3'; reverse primer for pGBK7g/pGADT7g, 5'-TTTTCGTTTAAAACCTAAGGTTCAC-3', for pDEST32/22 plasmids, 5'-AGCGCACAACCTTGATGGAAGAC-3'.

Retests. Interaction pairs were systematically retested twice as quadruplicates. Bait-and-prey pairs were rearrayed in complementary array format on separate 96-well plates including an additional prey strain that carries the empty prey plasmid (pDEST22 or pGAD7g, respectively) for each tested bait. This tested the bait for self-activation. Finally, diploids were pinned on Hi/Hi medium) by using the HIS3 reporter gene. The whole procedure was carried out on a Biomek FX laboratory workstation (Beckman Coulter) as described previously (12).

RESULTS AND DISCUSSION

General characteristics of the Dp-1 genome. The genome of phage Dp-1 is 56,506 bp in length (Fig. 1). Its G+C content is 40.3%, which is close to the 39% G+C content of S. pneumoniae, although two regions have lower G+C contents of 33.7% (between orf8 and orf9) and 36.7% (between orf35 and orf36), the latter being the start of a module transcribed in the opposite direction (see below). The region with the highest G+C content (48.5%) corresponds to the gene coding for the tail tape measure protein (orf52) and is located between 32,500 and 33,500 bp (Fig. 1).

Origin of replication. A putative origin (ori) was found near orf36 (positions 20,663 to 21,223 bp) using Ori-Finder (25). This region contains three partially conserved DnaA boxes (5'-TTATCCACA-3' consensus), with two mismatches per box (Fig. 1). The first DnaA box (position 20,793) starts immediately upstream of the potential ori. Interestingly, another potential ori was previously reported in the Dp-1 genome between the positions 56,348 and 56,464 bp, where two perfectly identical direct repeats were found (5'-TTATCCACAAGTG-3', orf39), which Ori-Finder fails to identify a potential ori in this region.

ORFs. The Dp-1 genome contains 72 ORFs, of which orf19, orf21, orf30, orf39, and orf62 were not previously identified in the partial Dp-1 genome description available (26), while five other ORFs (previously designated orf5, orf28, orf35, orf36, and orf40) presented in this partial description could not be confirmed here using our bioinformatics criteria. Among the identified 72 ORFs, 63 are transcribed from one strand, while nine ORFs (orf27 to orf35) are transcribed in the opposite direction and from the complementary strand (Fig. 1; see Table SA in the supplemental material). With the exception of orf46 and orf72, an RBS was found for each of the putative genes of Dp-1 but only 21 of these RBS presented a perfectly conserved Shine-Dalgarno consensus sequence. The average distance between the start codon and the end of the RBS was 8 nucleo-
tides. Out of 72 putative genes, 13 had an alternative starting codon closer to the RBS than an ATG starting codon (4 with GTG, 6 with TTG, 3 with ATT), while TAA was the most used stop codon (TAA, 49%; TAG, 31%; TGA, 20%). Overall, 42% of Dp-1 genes are overlapping. The coding capacity of the Dp-1 genome was estimated at 94%, which is similar to that of phage Cp-1 (60).

From the 72 ORFs, 39 (54.2%) were assigned a putative function based on homology to known proteins and 30 others have homologs in databases but with no assigned function, while 3 ORFs (orf8, orf9, orf20) code for novel proteins (Fig. 1 and Table 1; see Table SA in the supplemental material). The majority of the Dp-1 genes’ homologs are found in other phage genomes (34 of them), but a significant number of genes (27 of them) are also found in bacterial genomes. However, given that some of the bacterial homologous sequences may belong to nonidentified prophages, the overall contribution of phage-like genes to Dp-1 genome is probably higher. Among the bacteria, Clostridiales appear to be the major contributors to Dp-1 genomic makeup, with 8 BLAST hits.

**Transcriptional units.** We have identified 12 putative promoters in the genome of Dp-1. Each of them possesses a variable ~35 box (TTGACA consensus) and a conserved ~10 (TATAAT) sequence, while the distance between these two elements is variable among the promoters. Other less conserved promoters might be present in the genome of Dp-1, but in the absence of ~35 or ~10 consensus sequences they were not detected here. Promoters were found throughout the genome (Fig. 1; see Table SB in the supplemental material). The only perfectly conserved ~35 box was found in the putative promoter sequence located upstream of the genes involved in the initiation of the replisome formation, while the least conserved promoters were associated with genes coding for phage tail proteins. These differences among putative promoter sequences may reflect the timing and/or the intensity of transcription. We have also identified 7 putative terminator hairpins; the average free energy (∆G) of these secondary structures was 10 kcal/mol (see Table SB in the supplemental material). Based on the identified protein functions (see below) and these putative promoters and terminators, the genome of phage Dp-1 was divided into 8 functional clusters (Fig. 1 and Table SB).

**Queuosine synthesis gene cluster in phage Dp-1 genome.** orf1 to orf7 encode proteins predicted to be involved in queuosine biosynthesis, transport of queuosine precursor, and queuosine-tRNA modification (Fig. 1 and Table 1; see Table SC in the supplemental material). Queuosine is a hypermodified nucleoside derivative of guanosine (7-deazaguanosine) occupying the wobble position (position 34) of the tRNAs specific for Asp, Asn, His, or Tyr. The enzymes involved in synthesis and incorporation of queuosine are markedly different between eukaryotes and prokaryotes (64). While prokaryotes are able to synthesize queuosine from its precursors, eukaryotes must acquire queuosine from their diet or their microbiota (100). The biosynthesis of queuosine in bacteria is still incompletely understood (38). The modification of tRNAs by the inclusion of queuosine has an impact on the specificity of transcription and codon recognition but might also intervene in regulation of cellular physiology and metabolism (100). Numerous phages encode genes involved in queuosine biosyn-

---

### Table 1. Functional annotation of gene products of bacteriophage Dp-1

<table>
<thead>
<tr>
<th>ORF</th>
<th>Position (bp)</th>
<th>Putative function or domain</th>
<th>Predicted size (aa)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>orf1</td>
<td>131-652</td>
<td>Queuosine biosynthesis protein QueF</td>
<td>173</td>
</tr>
<tr>
<td>orf2</td>
<td>662-1348</td>
<td>Queuosine biosynthesis protein QueC</td>
<td>228</td>
</tr>
<tr>
<td>orf3</td>
<td>1350-1871</td>
<td>Queuosine biosynthesis protein QueD</td>
<td>173</td>
</tr>
<tr>
<td>orf4</td>
<td>1864-2658</td>
<td>Queuosine biosynthesis protein QueE</td>
<td>264</td>
</tr>
<tr>
<td>orf5</td>
<td>2534-3295</td>
<td>Tetrahydrofolate biosynthesis FolE</td>
<td>253</td>
</tr>
<tr>
<td>orf6</td>
<td>3306-3803</td>
<td>Queuosine biosynthesis intermediate transporter QueT</td>
<td>165</td>
</tr>
<tr>
<td>orf7</td>
<td>3793-4728</td>
<td>Queuosine tRNA-ribosyltransferase</td>
<td>311</td>
</tr>
<tr>
<td>orf10</td>
<td>5346-6419</td>
<td>DNA polymerase III subunit β DnaN</td>
<td>357</td>
</tr>
<tr>
<td>orf11</td>
<td>6419-7195</td>
<td>RecB exonuclease</td>
<td>258</td>
</tr>
<tr>
<td>orf12</td>
<td>7192-7683</td>
<td>RecA Holliday junction-specific endonuclease</td>
<td>163</td>
</tr>
<tr>
<td>orf14</td>
<td>8208-8699</td>
<td>dUTPase</td>
<td>163</td>
</tr>
<tr>
<td>orf15</td>
<td>8699-9859</td>
<td>RecA</td>
<td>386</td>
</tr>
<tr>
<td>orf16</td>
<td>9805-10218</td>
<td>NAD-dependent DNA ligase</td>
<td>137</td>
</tr>
<tr>
<td>orf17</td>
<td>10215-11240</td>
<td>DNA polymerase III, clamp loader complex, gamma/tau/ delta subunit</td>
<td>341</td>
</tr>
<tr>
<td>orf18</td>
<td>11242-12081</td>
<td>DNA polymerase III gamma/tau subunit</td>
<td>279</td>
</tr>
<tr>
<td>orf19</td>
<td>12074-12967</td>
<td>DNA polymerase III subunit delta</td>
<td>297</td>
</tr>
<tr>
<td>orf21</td>
<td>13160-14404</td>
<td>Metal-dependent phosphohydrolase HD</td>
<td>414</td>
</tr>
<tr>
<td>orf24</td>
<td>15081-15476</td>
<td>RNA polymerase, sigma 28 subunit</td>
<td>131</td>
</tr>
<tr>
<td>orf35</td>
<td>20411-20904</td>
<td>RNA 2-phosphotransferase, Ttp1/KptA family</td>
<td>105</td>
</tr>
<tr>
<td>orf37</td>
<td>22230-23621</td>
<td>Terminase, large subunit</td>
<td>463</td>
</tr>
<tr>
<td>orf38</td>
<td>23674-25434</td>
<td>Portal protein</td>
<td>386</td>
</tr>
<tr>
<td>orf40</td>
<td>25743-26738</td>
<td>Minor structural protein</td>
<td>331</td>
</tr>
<tr>
<td>orf41</td>
<td>26943-27611</td>
<td>Minor capsid protein</td>
<td>222</td>
</tr>
<tr>
<td>orf42</td>
<td>27267-28004</td>
<td>Structural protein</td>
<td>126</td>
</tr>
<tr>
<td>orf43</td>
<td>28017-29096</td>
<td>Structural protein</td>
<td>360</td>
</tr>
<tr>
<td>orf45</td>
<td>29451-29768</td>
<td>Structural protein</td>
<td>106</td>
</tr>
<tr>
<td>orf49</td>
<td>30896-31675</td>
<td>Structural Listeria Bacteroides repeat domain protein.</td>
<td>260</td>
</tr>
<tr>
<td>orf52</td>
<td>32386-35835</td>
<td>Tail length tape measure protein</td>
<td>1150</td>
</tr>
<tr>
<td>orf53</td>
<td>35847-36686</td>
<td>Tail protein</td>
<td>279</td>
</tr>
<tr>
<td>orf54</td>
<td>36698-40390</td>
<td>Receptor binding protein</td>
<td>1230</td>
</tr>
<tr>
<td>orf55</td>
<td>40401-42440</td>
<td>Minor structural protein</td>
<td>678</td>
</tr>
<tr>
<td>orf57</td>
<td>42774-43202</td>
<td>Signal peptide lipoprotein</td>
<td>142</td>
</tr>
<tr>
<td>orf58</td>
<td>43189-43413</td>
<td>Holin</td>
<td>74</td>
</tr>
<tr>
<td>orf59</td>
<td>43413-44303</td>
<td>Endolysin</td>
<td>296</td>
</tr>
<tr>
<td>orf60</td>
<td>44595-45299</td>
<td>S-layer protein membrane peptidase</td>
<td>234</td>
</tr>
<tr>
<td>orf61</td>
<td>45350-46987</td>
<td>Serine/threonine protein kinase, putative DNA binding protein. DNA/RNA helicase SNF2</td>
<td>545</td>
</tr>
<tr>
<td>orf66</td>
<td>48718-49362</td>
<td>DNA replication protein (DnaC)</td>
<td>214</td>
</tr>
<tr>
<td>orf67</td>
<td>49624-50961</td>
<td>DNA replicative helicase (DnaB)</td>
<td>445</td>
</tr>
<tr>
<td>orf68</td>
<td>50955-51974</td>
<td>DNA primase (DnaG)</td>
<td>339</td>
</tr>
<tr>
<td>orf69</td>
<td>52033-52647</td>
<td>RNA polymerase sigma factor</td>
<td>204</td>
</tr>
<tr>
<td>orf71</td>
<td>53538-55877</td>
<td>DNA polymerase I</td>
<td>779</td>
</tr>
<tr>
<td>orf72</td>
<td>55855-56388</td>
<td>Signal peptide membrane protein</td>
<td>177</td>
</tr>
</tbody>
</table>

* aa, amino acids.
thesis (see Table SC in the supplemental material), but none of these phages seem to have a complete queuosine biosynthesis cluster. The queuosine cluster of Dp-1 is apparently the most complete, even if it lacks the gene encoding QueA (see below). The Dp-1 genome contains the genes coding for the transporter QueT (orf6), as well as QueF (orf1), QueE (orf2), QueC (orf3), and QueD (orf4), which are needed for the synthesis of the queuosine precursors preQ0 and preQ1. It also encodes a gene (orf7) coding for a protein with similarity to the tRNA-guanine-transglycosylase needed for the replacement of guanine in the wobble position of tRNA with the queuosine precursor preQ1. As noted above, Dp-1 lacks the S-adenosylmethionine/tRNA ribosyltransferase-isomerase QueA, which is required for the transformation of the tRNA associated preQ0 into epoxy-queuosine (epoxyQ). However, QueA is encoded by the bacterial host S. pneumoniae, which lacks QueE and QueD homologs and has a tRNA-guanine-transglycosylase showing no similarity to the putative tRNA-guanine-transglycosylase homolog of Dp-1 (see Table SC).

It seems that phage Dp-1 and its host have complementary pathways for queuosine biosynthesis and queuosine-tRNA modification. No information is currently available on the possible function of the queuosine in S. pneumoniae. It is noteworthy that the queuosine synthesis cluster of Dp-1 is highly similar (E value = $3 \times 10^{-51}$) to the nonpathogenic Streptococcus thermophilus strain LMD-9 queuosine-coding genes. However, while it is tempting to speculate that Dp-1 uses queuosine-modified tRNAs to optimize its own translation, there is no experimental evidence for it. Moreover, interactions between Dp-1 structural proteins and queuosine synthesis proteins suggest alternative links between queuosine synthesis and the production of phage structural proteins (Fig. 2). One of the hypotheses might be that an excess of phage structural proteins, nonincorporated into virions and available to interact with proteins associated to queuosine synthesis, might limit the synthesis of queuosine, thereby reducing tRNA modification, and serve as a feedback signal to control phage assembly.

**DNA synthesis and repair clusters.** Two functional clusters involved in DNA replication and repair are located at positions 4,811 to 16,284 bp and 48,621 to 53,490 bp (Fig. 1, Table 1). Among other proteins, the first cluster encodes proteins forming the clamp-loader subunit of the DNA polymerase III complex (orf10, subunit; orf17, subunit; and orf18, subunit). The second replication cluster includes the putative genes encoding DNA replication proteins involved in the formation of the primosome (orf66, DnaC; orf67, DnaG).

The need for DNA recombination proteins Gp11 (RecB) and Gp15 (RecA) encoded in the first cluster might be due to a recombination-dependent replication model similar to phage T4 (61, 65). DNA repair might also be needed to counteract DNA damage induced by the strong oxidative stress produced by S. pneumoniae (74). The first replication cluster also contains a gene coding for a putative dUTPase (orf14) needed for the synthesis of the nucleotides required during the replication of phage DNA (29). Given that the Dp-1 genome does not appear to encode either a core DNA polymerase III enzyme, the primosome initiation protein DnaA, or a single-strand DNA binding protein, these parts of the replication apparatus might be provided by the host, while the specificity of the phage DNA replication might be ensured by the interaction of the host factors with the phage-encoded factors. Moreover, it is likely that phage proteins produced by both genomic clusters act in a cooperative fashion during Dp-1 DNA replication. It is tempting to speculate that the phage primosome is formed...
following the binding of the host DnaA to the replication origin by the addition of the other primosome-forming proteins encoded in the second replication cluster (DnaB, DnaC, and DnaG) and the host single-strand DNA binding protein. The core DNA polymerase III enzyme of the bacterial host would be redirected toward replication of the phage DNA through the use of the phage-encoded DNA polymerase III clamp (β subunit dimer) and clamp-loader (DNA polymerase III subunits γ-δ, Gp18; Δ,Gp19; and γ-δ, Gp17) thus completing the replisome formation. Notably, no phage homologs to the DNA polymerase III subunits β have been described so far (102).

Finally, the putative proteins from both DNA replication clusters are markedly different from pneumococcal proteins, showing instead homology to proteins encoded by the genes of an unnamed plasmid of *Eubacterium eligens* (see Tables SA and SB in the supplemental material). Because phages and plasmids are both vectors involved in gene transfer and contribute to the evolution of their prokaryotic hosts, it is not that surprising to detect similarity between phage and plasmid genes.

**DNA packaging cluster.** The DNA packaging cluster is localized between 12,998 and 23,621 bp and is interrupted by a cluster transcribed in the opposite direction between 16,281 and 20,431 bp (Fig. 1, Table 1). No small terminase subunit appears to be encoded in the genome of Dp-1. However, the packaging cluster contains a short gene (*orf23*) coding for a protein with distant homology to a part of the large terminase subunit gene (*orf37*) of *Streptococcus mitis* temperate phage EJ-1 as well as of a *S. pneumoniae* prophage (83). This genomic cluster also contains a gene (*orf21*) encoding a putative metal-dependent HD-phosphohydrolase possibly involved in the hydrolysis of the nucleic acid phosphodiester bonds (4, 107).

**Complementary strand-transcribed gene cluster.** As indicated above, the complementary strand-transcribed gene cluster is transcribed in the opposite direction, interrupts the DNA packaging cluster, and has a low G+C content, which suggests that this cluster might have been acquired via horizontal transfer (Fig. 1; see Table SB in the supplemental material). The genes of this cluster encode proteins with no defined function, the sole exception being *orf35*, possibly encoding a Tpt1/KptA family tRNA 2'-phosphotransferase. In *Saccharomyces cerevisiae*, Tpt1 is involved in the final step of tRNA splicing and catalyzes the transfer of the extra 2'-phosphate from the precursor-ligated tRNA to NAD (87). Our search for a *S. pneumoniae* Tpt1/KptA homolog yielded no results suggesting the absence of such function in this bacterium. Thus, the phosphate transfer from a 2'-phosphorylated RNA may play a role in Dp-1 replication.

**Morphogenesis.** The morphogenesis cluster can be divided into two groups of genes, one encoding capsid proteins (23,647 to 29,688 bp) and the other encoding tail proteins (35,782 to 43,107 bp). The capsid region comprises the genes encoding the portal protein (*orf38*), scaffolding protein (*orf41*), major capsid protein (MCP/*orf43*), two structural proteins identified by LC/MS-MS analysis of the purified virions (*orf42*, *orf45*), and a minor capsid protein (*orf40*) (Fig. 1; see Table SB in the supplemental material). Although Gp38 is different from portal proteins of other streptococcal phages, its distal region (amino acids 261 to 535) has similarities with the conserved domain of portal protein Gp6 of the well-characterized *Bacillus subtilis* phage SPP1 (PF05133) (70). The scaffolding protein Gp41 is homologous to the minor capsid protein Gp23 of *Staphylococcus aureus* phage 88 and to the conserved domain of the putative scaffolding protein Gp20 of *Lactobacillus* phage mv4. Distant similarities were also found with the putative scaffolding proteins of *S. pneumoniae* phages MM1 and EJ-1, suggesting a possible conservation among streptococcal phages.

The major capsid protein Gp43 was identified through its similarity with the conserved domain of the major capsid protein E of coliphage λ (PF03864). Although no *S. pneumoniae* phage homologs were found during PsIBLAST searches, several other streptococcal phage proteins present homology to Gp43, the closest being Gp31 of *S. thermophilus* temperate phage O1205. Gp42 does not present a significant similarity to any functionally characterized proteins, while Gp45 is homologous to several proteins of unknown function from different phages of *Clostridiales*. Gp40 has homology to the conserved domain of the minor head protein (F-like) of phage Mu (PF04233), as well as to a protein encoded by a *Clostridium botulinum* prophage.

The tail morphogenesis cluster comprises genes coding for a structural protein with a *Listeria-Bacteroides* domain (*orf49*), the tape measure protein (TMP) (*orf52*), a major tail protein (*orf53*), the receptor-binding protein (*orf54*), and a minor structural tail protein (*orf55*). Gp49 has homology to a protein from *Clostridium tetani* prophage 88 but was identified through its similarity to the distal part of *Listeria monocytogenes* internalins (with the PFAM conserved domain Flg-new), which is associated with cell wall surface proteins. The gene encoding Gp52 has the highest G+C content in the genome of phage Dp-1. While Gp52 presents homology to other TMPs of *S. pneumoniae* prophages, the best match is with a product encoded by a prophage-like gene from *Enterococcus faecium*. Two conserved domains were noticed in Gp52, one proximal (TIGR01760, amino acids 57 to 402) and one distal (COG5412, amino acids 370 to 1083).

Gp53 contains a conserved *Siphophila* family domain (PF05709) but exhibits no homology to any specific phage tail protein. Gp54 is likely the receptor-binding protein (RBP) (TIGR01665) involved in host recognition. It contains several GXY repeats and a collagen binding domain (CDD11035) that likely allows for the binding to the choline associated with the receptor on the cell surface of *S. pneumoniae* (55). Compared to other RBP from *S. pneumoniae* phages, a stretch of 9 amino acids is missing (between amino acids 733 and 734) in the region of collagen-binding motif of Gp54, thus shortening this domain by 3 GXY repeats.

Finally, Gp55 is a protein found in the Dp-1 virion structure by LC/MS-MS. When searched for homology by PsIBLAST, the first hit was with the Gp55 sequence of Dp-1 submitted to GenBank in 1997 (Z93946.1) (90). However, the latter sequence appears to contain a sequencing error resulting in a stop codon and a truncated ORF as well as a flanking short ORF covering the remaining sequence of Gp55. When these errors are corrected, Gp55 appears highly similar to other streptococcal phage sequences, one of them being the minor tail protein Gp49 of the pneumococcus phage MM1. All these
Gp55-like proteins contain a domain of unknown function (DUF859) commonly found in phages of the family Siphoviridae.

**Lysis cluster.** The lysis cluster is located between 43,107 and 44,468 bp and is flanked by a putative promoter and terminator. This region contains two genes, namely, orf58, coding for the holin (Dph), and orf59, coding for the N-acetylmuramoyl-L-alanine amidase endolysin (Pal). Many homologs to Gp58/Dph were found among *Enterococcus* phages but none with streptococcal phages. As indicated previously, the endolysin Pal has been previously characterized and successfully used to lyse pneumococcus cells *in vitro* and treat *S. pneumoniae* infection in murine models (39, 52, 53, 80).

**Sigma factors.** Dp-1 genome encodes two putative sigma-like factors, namely, Gp24 and Gp69. These proteins are markedly different from *S. pneumoniae* sigma factors and do not share similarity to each other. As Dp-1 does not appear to encode a RNA polymerase, it likely uses the host enzyme for transcription and the sigma factors may redirect the pneumococcus RNA polymerase to the phage promoters. Gp24 contains region 3 and 4 domains of sigma70 factors (PF04539 and PF04545). It also has a low level of similarity to sigma factor 28 and to two prophage-encoded proteins in two strains of *S. pneumoniae*. Gp69 is homologous to the sporulation SigK sigma factor of *Thermoanaerobacter pseudethanolicus* and contains two sigma factor domains, sigma70 regions 2 and 4 (PF04542 and PF04545). It was recently showed that a peptide derived from a phage anti-sigma factor is able to inhibit *S. aureus* transcription (18). Whether the sigma factors of phage Dp-1 have similar activities remains to be investigated.

**The binary Dp-1 interactome.** Protein-protein interactions can provide useful information on functional connections between proteins. We carried out two comprehensive Y2H screens, testing all pairwise interactions among Dp-1 proteins using two different Y2H vector systems. Our screens detected a total of 232 redundant interactions among the 72 deduced proteins of phage Dp-1. Of those, 99 were identified with the pDEST32/22 and 133 with the pGBK/GADT7g vector system. We retested all these interactions and reproduced 69 (69.7%) of the pDEST32/22 pairs and 109 (82.0%) in pGBK/GADT7g. Overall, we found 156 unique, nonredundant protein-protein interactions among the restet set (Fig. 2; see Tables SD and SE in the supplemental material). On average each Dp-1 protein binds 2.2 interaction partners and 57 proteins (out of the 72) exhibited at least one interaction. Thus, 79.2% of the deduced proteome participates in the interaction network described herein.

The pDEST32/22 system identified 49 (31.4%) and the pGBK7g/pGADT7g system identified 92 (59.0%) unique interactions of the 156 nonredundant interactions. Only 15 of them overlapped between both vector systems (9.6%). This number is in agreement with a previous study, although the overlap here is about three times higher (23). With the pGBK7g/pGADT7g expression system we obtained nearly twice as many unique interactions (see Fig. S1 and Table SD in the supplemental material). The pDEST32/22 interaction network consisted of 51 nodes connected by 64 edges. This corresponds to an average node degree of 1.26. In contrast, the pGBK7g/GADT7g network contains 46 proteins connected by 107 interactions. Although the pDEST system detected fewer interactions, it included five additional proteins in the network and thus may be more specific.

We also estimated the interaction strength using various concentrations of 3-AT, a competitive inhibitor of the reporter gene product HIS3 (35). We determined for each interaction pair the highest 3-AT concentration on which yeast growth was observable, and we grouped the interactions as weak, medium, or strong. Overall, the pDEST32/22 system resulted in stronger interactions (40.6%) compared to pGBK/GADT7g (16.5%) (see Fig. S2 in the supplemental material).

**Dp-1 genomic clusters and protein-protein interactions.** We wondered whether proteins primarily interact with functionally related proteins (Fig. 3). Hence, we counted the number of interactions that occur between proteins, in which their respective gene is located within the same functional cluster. Indeed, the majority of interactions (64.6%) occurred between proteins belonging to the same cluster. However, 35.4% of the interactions were between proteins from different functional clusters, indicating that there is an active cross talk. For example, structural proteins showed 27 intrafunctional interactions while proteins involved in DNA replication underwent 29 intrafunctional interactions. These two groups cross-communicate via 27 additional interactions, indicating a close connection between DNA replication and virion assembly.

**Model of Dp-1 virion.** In order to develop a model of the Dp-1 virion, we integrated genome annotation, binary interactions, and LC/MS-MS data (Fig. 4). By using LC/MS-MS we identified 8 proteins in purified Dp-1 particles, primarily known structural proteins, namely, Gp38 (portal), Gp42 (capsid-associated protein), Gp43 (MCP), Gp45, Gp49 (tail protein), Gp52 (tape measure protein), Gp54 (RBP), and Gp55 (tail protein) (labeled in red in Fig. 4). Three other proteins (Gp40, Gp41, and Gp53) could be linked by functional annotation as well as by binary protein-protein interactions and might function as morphogenetic factors such as chaperones dissociating from the mature virion (92). For example, the MCP Gp43 interacts with Gp41, which is a homolog of the minor structural protein Gp20 of *Staphylococcus* phage φETA.

Gp41 could be a capsid scaffolding protein, since it interacts specifically with Gp43, but does not appear to be a structural component as determined by LC/MS-MS. Ten other gene products were linked to the virion structure by Y2H interactions (proteins labeled in blue in Fig. 4). Several predicted structural proteins were not detected by LC/MS-MS but found to interact in Y2H screens as homomers, such as Gp56. Another protein, Gp33, links head and tail proteins (Fig. 4A). Finally, we found Gp42, Gp45, and Gp49 in the virion structure by LC/MS-MS but could not find interaction partners. As a consequence, we could associate 21 proteins encoded by the morphological/structural module, including several hypothetical proteins.

**False negatives.** We most likely could not detect all protein-protein interactions. Any particular Y2H system is estimated to detect about 25% of all interactions (14). We used two different Y2H vector systems to increase the screen sensitivity, as shown by others (23). We also added MS data to increase the interaction coverage. Still, we could not observe or infer interactions in Dp-1 that were previously detected in other phages. In Fig. 4C, we summarize known interactions found in other phages that might have escaped detection in our study.
Unexpected interactions with virion components. The eight proteins identified by LC/MS-MS were expected, since they are homologous to known virion proteins. In addition, we found 12 nonstructural gene products that interact with the structural core (Fig. 4C). For example, we found interactions between enzymes involved in queuosine metabolism (QueD, QueE, QueF, QueT) with proteins involved in virion assembly. As suggested previously, these interactions may function in vivo as feedback signals for protein translation. Proteins such as Gp44, Gp33, or Gp55 could block the Que enzymes when a critical protein level is reached, saving energy spent on translation and queuosine biosynthesis. Ten additional interactions were also detected that involved the putative RNA polymerase sigma factor Gp69.

As indicated previously, we identified interactions between structural and replication proteins. For example, the DNA polymerase I (Gp71), DNA polymerase III subunit β (Gp10), DNA ligase (Gp16), and DNA primase (Gp68) were found to interact with proteins likely involved in DNA packaging (Fig. 4C). In fact, the observed interactions link terminase assembly, DNA packaging, and replication. Note that in coliphage T4 the packaging process involves only phage proteins. Moreover, Holliday structure intermediates that occur during T4 DNA replication have a negative effect on DNA packaging but can be resolved by a T4-encoded endonuclease that binds the portal protein (41, 57). The fact that DNA replication/repair enzymes like T4 DNA ligase are needed in vivo for efficient packaging suggests a close association of DNA packaging and DNA replication/repair processes (106). In phage Dp-1, the interaction of Gp23 with Gp16 (DNA ligase) suggests that ligase could repair DNA nicks during packaging. Recruitment of Dp-1 Gp71 (DNA polymerase I) via Gp22/Gp23 interaction could also initiate DNA repair or primase primer displacement in parallel to packaging. Interestingly, the Gp23 of Dp-1 also interacts with the DNA polymerase III β-clamp subunit (Gp10), while in T4 the large terminase subunit was shown to interact with the β-clamp subunit (58). This suggests a role of the β-clamp/DNA polymerase III in DNA packaging in vivo, which cannot be simulated in the pairwise interaction tests carried out here.

DNA packaging and terminase. In many tailed phages which use a headful packaging mechanism, efficient DNA encapsidation requires an active portal protein embedded in the procapsid portal vertex as well as an interacting terminase complex (21, 37). Although we could not detect any interaction of the Dp-1 terminase (Gp37) with the portal protein, we discovered several interactions (Fig. 4C) that link terminase-associated proteins to capsid components: Gp39, Gp23, and Gp33 bind to...
Gp43 (major capsid protein), Gp22 to Gp41 (minor capsid component), and Gp33 to Gp44 (predicted lipoprotein). These interactions could mediate the terminase assembly in a cooperative manner with the portal vertex. The terminase-associated Gp23 has a low level of similarity with large terminase subunits from other phages and may act as a hub by binding virion-associated proteins (Gp49, Gp22, Gp48, Gp39). Thus, Gp23 could be important for recruiting additional protein factors that stimulate the packaging process to the terminase. It is interesting that the Gp23 sequence does not contain a helix-turn-helix motif, which is a key feature for the small terminase subunit in many phages and important for cooperative DNA binding (17) and DNA packaging. Thus, Gp23 cannot be viewed as a canonical small terminase subunit, but it might act more as an adaptor than a protein directly binding the DNA substrate. However, the hypothetical protein Gp39 does contain a predicted DNA-binding motif (DNA polymerase alpha zinc finger, PF08996) and associates via Gp36 with the large terminase subunit. Gp39 could be the key component of the terminase complex by binding the DNA substrate. Notably, in the genome of other long-tailed pneumococcal (MM1) or S. mitis (EJ-1) phages, a canonical small terminase subunit appears also to be missing (68, 83).

Capsid proteins and lipids. It was previously reported that phage Dp-1 particles contain lipids (56). This rare combination has not been reported for any other phages of the order Caudovirales, but it was found for Pseudoalteromonas phage PM2 (family Corticoviridae), for Pseudomonas phage phi6 (Cystoviridae), and for enterobacterial phages of the PRD1 group (Tectiviridae) (6, 24, 99). In the nontailed double-stranded RNA phage phi6, the lipids surround the capsid and contain additional envelope proteins and lipoproteins (5). Members of the Corticoviridae contain a lipid vesicle between the outer and inner layer of their capsid, while Tectiviridae contain a lipid bilayer inside their capsid. The exact location of the lipids in Dp-1 is unknown. As described above, the hypothetical protein Gp44 of Dp-1 carries an N-terminal signal peptide and is predicted to be a lipoprotein by its similarity to a lipoprotein domain (PROSITE PS51257). We found it to interact with the major capsid protein Gp43. It is tempting to speculate that
Gp44 is inserted via its N-terminal signal peptide into the cell membrane during intracellular phage replication. Its interaction with the major capsid protein may indicate that the phage head assembly is localized to the membrane. Interestingly, phage T4 head assembly is localized to the inner side of the cytoplasmic membrane, at which it forms a membrane-spanning initiator complex consisting of T4 portal protein and Gp40. After the release of the phage head into the cytoplasm, Gp40 is no longer interacting with assembled capsids (36). An analogous situation might be present in Dp-1. The lipoprotein Gp44 was not found by LC/MS-MS, and a membrane-dependent procapsid assembly could be initiated by the Gp43/Gp44 interaction rather than by the portal protein.

**Tail.** LC/MS-MS analysis confirmed Gp52 (tail length tape measure protein), Gp54 (RBP), and Gp55 (minor structural protein) as structural proteins. The last of these carries a Pfam signature (DUF859), which is found in at least 44 protein sequences, mainly in *Streptococcus* phages. DUF859 proteins are annotated as “structural” or “minor tail proteins.” Protein Gp55 is probably the major component of the Dp-1 tail, and its homomic interaction confirms its ability to polymerize. In addition, Gp55 interacts with the receptor-binding protein Gp54 (Fig. 4C). No binary contact was found with the TMP protein that acts as a ruler of the tail length, possibly because it requires cooperative binding to assembled Gp53 and/or other chaperone proteins (92).

**Conclusions.** The diversity of protein homologs in phage Dp-1 highlights the rich and diverse genetic information content of the virosphere (16, 32). The fact that the majority of Dp-1 genes are homologous to those of other phage (either virulent or temperate) sequences illustrates the importance of gene transfer and recombination among phages (30). Clearly, pneumococcal phages have the same genome plasticity as their hosts. Because of the similarities with prophages, it is very likely that virulent phages also play a role in the evolution and the diversification of the clinical isolates.

This host diversification is illustrated by over 90 different capsular serotypes (11, 13). To our knowledge, no pneumococcal phage, including Dp-1, is currently able to efficiently infect capsulated *S. pneumoniae* strains (9). Still, it is assumed that Dp-1 must be able to infect its host cells in its ecological niche of the human upper respiratory tract. Phages infecting capsulated bacteria often encode specific glycosyl hydrolases (depolymerases) targeting the capsular polysaccharide (67, 89). Based on our bioinformatic analyses, a gene coding for a capsule depolymerase apparently does not exist in Dp-1. However, the putative tail protein Gp53, which contains a *Sipho* tail superfamily domain (PF05709), has a distant homology to a glycosyl transferase. A comprehensive understanding of the phage-host interactions in pneumococci may lead to the isolation of phages infecting such capsular strains or to the identification of physiological conditions favoring the phage infection process.

This study also describes one of the first genomes whose annotation has been aided by binary protein interaction data. Our intraviral interactome also led to a plethora of new hypotheses on the protein network needed to complete a successful phage lytic cycle. The biological relevance of these interactions still remains to be investigated, but comparison with other intraviral interactomes will likely lead to a better understanding of general and specific strategies used by phages to maximize their replication. Moreover, the protein-protein interactions found in this study are also unlikely to occur in isolation. Phage proteins do not only interact with each other but also with host proteins, and these interactions are influenced by the physiological condition of the host. Some interactions were probably not detected because it may not be possible to identify them outside the host. Our current phage-phage interactome, while being undoubtedly incomplete, still provides several starting points for future studies. Our ranking based on interaction strength and reproducibility helps with the prioritization of such studies. The next level of analysis will require the screening of phage proteins against libraries of host proteins. This global phage-host analysis will certainly also uncover novel aspects of phage biology in the context of the host cell as well as provide a better understanding of the phage-host interactions, possibly leading to novel therapeutic approaches for this major human pathogen.

**ACKNOWLEDGMENTS**

We thank Denise Tremblay for discussion and electron micrographs. This project was supported by the Landesstiftung Baden-Württemberg (Germany), the Karlsruhe House of Young Scientists (R.H.), NIH grant RO1GM79710 (P.U.), and the Canadian Institutes of Health Research (Team Grant—Emerging: Novel Alternatives to Antibiotics) (M.O. and S.M.).

**REFERENCES**


