T4-Like Genome Organization of the *Escherichia coli* O157:H7 Lytic Phage AR1

Wei-Chao Liao,1 Wailap Victor Ng,1 I-Hsuan Lin,3 Wan-Jr Syu,4 Tze-Tze Liu,3 and Chuan-Hsiung Chang2,3,*

Department of Biotechnology and Laboratory Science in Medicine,1 Center for Systems and Synthetic Biology,2 Institute of Biomedical Informatics,3 Institute of Microbiology and Immunology,4 and Genome Research Center,5 National Yang-Ming University, Taipei, Taiwan

Received 15 November 2010/Accepted 4 April 2011

We report the genome organization and analysis of the first completely sequenced T4-like phage, AR1, of *Escherichia coli* O157:H7. Unlike most of the other sequenced phages of O157:H7, which belong to the temperate *Podoviridae* and *Siphoviridae* families, AR1 is a T4-like phage known to efficiently infect this pathogenic bacterial strain. The 167,435-bp AR1 genome is currently the largest among all the sequenced *E. coli* O157:H7 phages. It carries a total of 281 potential open reading frames (ORFs) and 10 putative tRNA genes. Of these, 126 predicted proteins could be classified into six viral orthologous group categories, with at least 18 proteins of the structural protein category having been detected by tandem mass spectrometry. Comparative genomic analysis of AR1 and four other completely sequenced T4-like genomes (RB32, RB69, T4, and JS98) indicated that they share a well-organized and highly conserved core genome, particularly in the regions encoding DNA replication and virion structural proteins. The major diverse features between these phages include the modules of distal tail fibers and the types and numbers of internal proteins, tRNA genes, and mobile elements. Codon usage analysis suggested that the presence of AR1-encoded tRNAs may be relevant to the codon usage of structural proteins. Furthermore, protein sequence analysis of AR1 gp37, a potential receptor binding protein, indicated that eight residues in the C terminus are unique to O157:H7 T4-like phages AR1 and PP01. These residues are known to be located in the T4 receptor recognition domain, and they may contribute to specificity for adsorption to the O157:H7 strain.

Bacteriophages (or phages) are the most abundant entities in the biosphere, and they play an important role in microbial evolution and pathogenesis (2). For instance, a number of phages have been isolated from the notorious *E. coli* strain O157:H7, which is an important and well-known human pathogen in uncooked and contaminated food (20) that may cause serious complications of bloody diarrhea, thrombotic thrombocytopenic purpura (TTP), and hemolytic-uremic syndrome (HUS) (58). The major pathogenic factor of *E. coli* O157:H7 is the Shiga toxin (Stx), encoded by Stx-producing temperate phages (24).

The genomes of 12 *E. coli* O157:H7 phages have been sequenced completely and published in the NCBI genome database. Most of them belong to the temperate *Podoviridae* and *Siphoviridae* phage families and carry Stx-like toxin genes. These include the converting phages Stx1 (72), Stx2-I, Stx2-II (71, 74), and Stx2-1717 (GenBank accession no. NC_011357; Y. Zhang et al., unpublished data); the enterobacterial phages 933W (64), VT2-Sakai (40, 47), Min27 (36, 77), and YYY-2008 (GenBank accession no. NC_011356; Zhang et al., unpublished data), which encode the Stx-like proteins; and the non-toxin-carrying phages phiV10 (61) and JK06 (GenBank accession no. NC_007291; J. Kagan et al., unpublished data). Genome comparisons showed that the phages with Stx-like toxin genes have closely related genome sequences (40, 71, 72). Two other sequenced lytic phages, RV5 and WV8, are the only two known *E. coli* O157:H7 lytic phages which belong to the *Myoviridae* phage family (53, 54). However, these two phages are dissimilar in genome size, genome features, and DNA sequence in comparison with phages of the other genera of the *Myoviridae* family, such as the T4-like phages (85).

AR1, a lytic phage that can efficiently infect the *E. coli* O157:H7 strain, was identified from the stool of cows by Ronder and Cliver in 1990 (70). The viral particle contains an isometric head and a contractile tail of 103 by 74 and 116 by 16 nm, respectively (17). It has a typical T4-like virion morphology under a transmission electron microscope. AR1 has a wide host range (17, 89) and can efficiently infect and lyse both *Shigella dysenteriae* strains. In addition, AR1 can also lyse other *Escherichia coli* serovars Choleraesuis and Enteritidis (17, 89). Host range receptor assays (17) showed that AR1 uses the R1, R3, and R4 core oligosaccharide (OS) types of lipopolysaccharide (LPS), but not the R2 or K-12 type, as cellular receptors (17). A previous study suggested that the putative AR1 gp37 receptor recognition protein may recognize and bind to the OmpC receptor because an *E. coli* O157:H7 OmpC disruption mutation rendered the cell resistant to AR1 infection (90). Similar to the case with T4 and other T4-like phages, restric-
tion enzyme digestion experiments suggested that the AR1 genome is highly modified to protect it from restriction enzyme digestion by the host defense system (89). Partial sequencing data (~10 kb) of AR1 indicated that AR1 has several structural genes that are highly similar to T4 genes (17, 37, 89).

The T4 phage family is one of the best-characterized groups of E. coli phages (45, 81). Although more than 200 T4-like phages have been examined (4), only a limited number of these phage genomes have been sequenced completely (T4-like genome database [http://phage.bioc.tulane.edu], Tulane University Health Sciences Center, New Orleans, LA). Most of the known T4-like phages specifically infect certain strains of E. coli or some other enterobacteria, but several T4-like phages can propagate in phylogenetically more distant bacteria, such as Aeromonas, Acinetobacter, Vibrio, and Cyanobacteria (4, 5, 8, 44, 62). Previous cross-genome comparisons of T4 and other T4-like phages (41, 44, 55, 62, 78) provided many interesting findings in the field of phage biology. It appears that this family of phages shares a common core genome from an ancestral sequence encoding the DNA replication modules, virion structural proteins, and some conserved hypothetical proteins (7, 8, 62). Furthermore, some of the proteins with unknown functions have been found to be conserved in T4-like phage genomes (8). In addition, regions showing high diversity (hyperplastic regions [HPRs]) in T4-like genomes are also as important as the conserved regions. These regions contain several predicted functional genes, such as the internal protein (IP) genes, genes encoding distal tail fibers of receptor recognition proteins, DNA modification genes, and various numbers and types of phage-carried tRNA genes (8), which may give the phages unique biochemical or physiological properties (8, 12). Previous analysis also speculated that some of the minor differences between these phages may be related to their adaptations to different host ranges (8, 12).

In addition to AR1, four T4-like phages (CEV1 [66], e112 [56], PP01 [49, 56], and ECML-134 [1]) have also been discovered to efficiently infect and lyse the E. coli O157:H7 strain BE. However, none of the E. coli O157:H7 T4-like phages have been sequenced completely. Here we report the genome organization and results of proteomic and comparative genomic analyses of AR1 and its related T4-like phages. The whole-genome analysis of AR1 may shed light on both host specificity and genome organizations of the T4-like phages.

MATERIALS AND METHODS

Bacterial and phage strains. E. coli O157:H7 strain 1266 was originally isolated in Japan and was kindly provided by H. Watanabe, National Institutes of Health (NIH) (23). E. coli K-12 strain MG1655 (ATCC 47076) and E. coli B[2] (ATCC 11303) were obtained from the American Type Culture Collection (ATCC). All of the E. coli strains were cultured at 37°C in Luria-Bertani broth with agitation (180 rpm). AR1 phage were propagated in E. coli O157:H7 strain 1266 as described previously (70, 89). E. coli phages T4 and RB69, obtained from J. D. Karam (Tulane University), were propagated in E. coli strain B[2] (92).

Preparation of phage genomic DNA. Phage stocks of AR1, T4, and RB69 were prepared according to a previously described method (70, 89). Fresh phage particles of AR1 precipitated with 20% (wt/vol) polyethylene glycol 8000 (PEG 8000) and 2.5 mol/liter NaCl were resuspended in 1 ml STE buffer (10 mM Tris-Cl, 100 mM NaCl, and 1 mM EDTA, pH 8.0). The viral suspensions were treated with 10 mg of RNase A (Sigma) and 100 U of DNase (Roche) at 37°C for 1 h to remove the host genomic DNA and RNA. The mixtures were then adjusted to a final concentration of 1% sodium dodecyl sulfate (SDS) and digested with 20 U of proteinase K (Invitrogen) for 1 h at 37°C. The AR1 genomic DNA was extracted using a phenol-chloroform (1:1) solution and then ethanol precipitated according to a previously described procedure (89). The viral genomic DNA was examined by PCR to ensure that there was no contamination of host genomic DNA.

Genome sequencing. Both automated ABI 3730 capillary DNA analyzer (Applied Biosystems, Foster City, CA) (14) and Genome Sequencer 20 (GS20; 454 Life Science Corporation, Branford, CT) (42) systems were used to analyze the AR1 whole-genome sequence. Briefly, for fluorescence DNA sequencing, the DNA samples were fragmented by a Hydroshear machine (GeneMachines, San Carlos, CA), and then 2- to 3-kb fragments were collected, ligated with pUC18 vector, and transformed into E. coli DH5a. The sequencing templates were prepared by use of a Templiphi amplification kit (GE Healthcare, Piscataway, NJ) directly with cultured cells, sequenced with a BigDye Terminator v3.1 cycle sequencing kit (Applied Biosystems), and analyzed on an ABI 3730xl DNA analyzer (Applied Biosystems). In order to increase the sequence quality, a Genome Sequencer 20 system was also used in this study. The DNA samples were nebulized to 500- to 800-bp fragments and then ligated to the adaptor to generate a library of random single-stranded template DNA fragments (ssTDNA). The ssTDNA library was immobilized onto beads and amplified by emulsion PCR technology. Following amplification, the DNA-carrying beads were loaded on the GS20 sequencer (Roche Applied Science) for one run of pyrosequencing. Approximately 75,230 reads (45-fold coverage) from the GS20 sequencer, and sequences were assembled by a Newbler de novo shotgun sequence assembler. The resulting contigs were assembled together with the ABI sequencing reads by Phred/Phrap/Consed software.

Annotation of AR1 genome sequence. A customized gene prediction pipeline was used to predict the open reading frames (ORFs) of the AR1 phage genome. The results of the genome annotations were compared to the GeneMark.hmmer (version 2.6) (39) and Zcurve_V (version 1.0) (19), with a minimal size of 30 amino acids. The Blastp program (version 2.2.15) (6) from NCBI was used to search for sequence similarity of the predicted ORFs against the NCBI nonredundant protein, KEGG gene, and UniProt/TrEMBL protein databases. Blastp matches sequences with a minimum of 30% identity; sequences with a subject sequence coverage of >70% were considered possible homologs, and their functions were assigned to the query AR1 ORFs. Coding sequences with no Blastp matches were designated hypothetical proteins, and those that matched hypothetical or putative proteins in the database were designated conserved hypothetical proteins. Viral clusters of orthologous groups (viral COG or VOG; http://www.ncbi.nlm.nih.gov/genomes/VIRUSES/vog.html) were assigned to the predicted AR1 proteins based on sequence similarity. Seven T4-like phage genomes (RB49, RB52, RB69, T4, JS98, 44RR, and RB43) were also analyzed with the same annotation and VOG assignment pipeline for optimal comparative genomic analysis. In order to increase the functional assignments of AR1 predicted proteins, protein domain information was also considered. The protein sequences of conserved hypothetical and hypothetical proteins were searched for the presence of known domains by use of InterProScan software and the InterPro database, with an E value cutoff of <0.001 (91). Viral proteomic and structural analyses. Viral proteomic analyses of AR1 were performed using TCEP-HCl (Pierce, Rockford, IL) and digested with trypsin at 37°C overnight, similar to a previously described method (16). Trypsin-digested peptides were purified using a PepClean C18 spin column (Pierce, Rockford, IL) and analyzed by LTQ Orbitrap hybrid mass spectrometers (Thermo Scientific, San Jose, CA). Mass spectrometry (MS) raw data in Thermo XCalibur (version 2.0) binary format were converted to the mzXML open data format by using a modified version of the ReAdW program (59). The output files were then analyzed with the target-decoy search strategy as described by Elias and Gygi (13) to evaluate the false discovery rate (FDR). The SEQUEST (version 27) algorithm (87) was applied to search the MS/MS spectra against a database containing the forward and reverse (decoy) sequence coverage of proteins in the GenBank nonredundant protein database and against the AR1 protein sequences. Search parameters included a peptide mass tolerance of ±2.1 Da and possible oxidation of the methionine (+16.0 Da) residue without specifying the cleaving agent for nonconstraint database matching. SEQUEST results were further processed using the Trans Proteomic Pipeline (26, 27, 52). An identification probability value of 0.9 (equivalent to an estimated 10% error rate) for both peptides (evaluated by PeptideProphet) and proteins (evaluated by ProteinProphet) was used as the cutoff. Only the proteins (P ≤ 0.9) with protein-specific unique peptides identified by tandem mass spectrometry were considered positive identifications, unless specified otherwise. Additional identification criteria included a maximum of two missed tryptic cleavages and precursor ions carrying up to five protons.

Comparative genomic analysis. The phage genome sequences downloaded from the NCBI, Tulane T4-like genome (http://phage.bioc.tulane.edu), and
TABLE 1. General genome features of AR1 and the other sequenced T4-like phages

<table>
<thead>
<tr>
<th>Phage</th>
<th>GenBank accession no.</th>
<th>Genome size (bp)</th>
<th>No. of predicted ORFs</th>
<th>GC%</th>
<th>Gene densitya</th>
<th>% Coding sequence</th>
<th>Avg gene product size (aa)b</th>
</tr>
</thead>
<tbody>
<tr>
<td>RB49</td>
<td>NC_005066</td>
<td>164,018</td>
<td>279</td>
<td>40.0</td>
<td>1.7</td>
<td>94</td>
<td>184</td>
</tr>
<tr>
<td>RB14</td>
<td>NC_012638</td>
<td>165,429</td>
<td>274</td>
<td>35.0</td>
<td>1.7</td>
<td>94</td>
<td>189</td>
</tr>
<tr>
<td>RB32</td>
<td>NC_008515</td>
<td>165,890</td>
<td>270</td>
<td>35.0</td>
<td>1.6</td>
<td>94</td>
<td>193</td>
</tr>
<tr>
<td>AR1</td>
<td>AP011113</td>
<td>167,435</td>
<td>281</td>
<td>35.3</td>
<td>1.7</td>
<td>95</td>
<td>198</td>
</tr>
<tr>
<td>RB69</td>
<td>NC_009428</td>
<td>167,560</td>
<td>273</td>
<td>37.0</td>
<td>1.6</td>
<td>93</td>
<td>190</td>
</tr>
<tr>
<td>RB51</td>
<td>NC_012635</td>
<td>167,394</td>
<td>273</td>
<td>35.0</td>
<td>1.6</td>
<td>94</td>
<td>193</td>
</tr>
<tr>
<td>T4</td>
<td>NC_000866</td>
<td>168,903</td>
<td>278</td>
<td>35.0</td>
<td>1.6</td>
<td>93</td>
<td>188</td>
</tr>
<tr>
<td>JS98</td>
<td>NC_010105</td>
<td>170,523</td>
<td>266</td>
<td>39.0</td>
<td>1.6</td>
<td>91</td>
<td>194</td>
</tr>
<tr>
<td>JS10</td>
<td>NC_012741</td>
<td>171,451</td>
<td>265</td>
<td>39.0</td>
<td>1.5</td>
<td>92</td>
<td>211</td>
</tr>
<tr>
<td>44RR</td>
<td>NC_005135</td>
<td>173,591</td>
<td>252</td>
<td>43.0</td>
<td>1.5</td>
<td>92</td>
<td>211</td>
</tr>
<tr>
<td>RB43</td>
<td>NC_007023</td>
<td>180,500</td>
<td>292</td>
<td>43.0</td>
<td>1.6</td>
<td>93</td>
<td>192</td>
</tr>
</tbody>
</table>

a All analyzed phages are members of the T4-type viruses, except for 44RR of the 44RR2.8t-type viruses, RB43 of the RB43-type viruses, and RB49 of the RB49-type viruses (34, 63). The host ranges of all T4-like phages are E. coli K-12 or B strains, except for AR1 and 44RR phages, which infect E. coli O157:H7 and A. salmonicida, respectively. All phages are lytic phages containing linear double-stranded DNA. General genome information was obtained from the NCBI, Tulane T4-like genome (http://phage.ggc.edu), and GGC T4-like genome (http://phage.ggc.edu) websites.

b Gene density = predicted gene number/genome size (bp).

c Average gene product size = total coding length (aa)/predicted gene number.

GCG T4-like genome (http://phage.ggc.edu) websites for genome comparisons are listed in Table 1. The Dotter program (version 3.0) was used to examine whole genome similarity, using a sliding window of 25 bp (76). Basic genome features of all phages, such as their host specificity, taxonomy, life cycle, and genome molecular type, were also collected from the NCBI database and the literature. GenSkew (http://genskew.csb.univie.ac.at/) was used to compare the sequence skew properties of the phage genomes. Potential tRNA genes in the genome sequences were identified using three tRNA prediction tools, namely, tRNAscan-SE (version 1.21) (38), ARAGORN (version 1.2) (33), and Rfam (version 8.1) (18). AR1 phage codon usage was analyzed with the CUSP and CAI programs of the EMBOSS, version 6.2.0, package (68). The putative proteins from the downloaded phage genomes and the proteins from AR1 were compared for sequence similarity by use of the Blastp program (cutoff for sequence identity, >30%; cutoff for sequence coverage, >70%; cutoff for E value, <1 × 10-10).

Nucleotide sequence accession numbers. The nucleotide sequence of phage AR1 has been deposited in the DDBJ database under accession number ABP11113. The GenBank accession numbers of the other phage genomes analyzed in this study (listed in Table 1) are as follows: RB49, NC_005066 (8, 11, 12); RB14, NC_012638 (J. M. Nolan et al., direct submission); RB32, NC_008515 (Nolan et al., unpublished data); RB69, NC_009428 (55, 62); RB51, NC_012635 (Nolan et al., direct submission); T4, NC_000866; JS98, NC_010105 (92); JS10, NC_012741 (10); 44RR, NC_005135 (55, 81); and RB43, NC_007023 (55, 52). The GenBank accession numbers for internal protein sequences are as follows: T4 IP1, NP_049739; T4 IP2, NP_049734; T4 IP3, NP_049735; RB14 IP1, NP_00285447; RB14 IP4, NP_00285446; RB14 IP5, NP_00285447; RB32 IP4, NP_080305; RB32 IP5, NP_080308; RB32 IP7, NP_080301; RB32 IP9, NP_080302; RB51 IP4, NP_00285408; RB51 IP5, NP_00285409; RB69 IP1, NP_061845; RB69 IP5, NP_061817; RB69 IP7, NP_061846; RB69 IP10, NP_061814; JS10 IP3, NP_00292246; JS10_112, NP_00292246; JS98 IP3, NP_00195924; and JS98_114, NP_00195924. The GenBank accession numbers for g37 protein sequences are as follows: AR1, BA053266; P001, AAK30164; T4, NP_049863; RB69, NP_061944; and JS98, NP_00195372. The GenBank accession numbers for OmPC protein sequences are as follows: E. coli K-12 MG1655, NP_416719; O157:H7 1266, AAF21761; and ATCC 43888, NP_288795.

RESULTS

AR1 is closely related to other sequenced T4-like phages. In previous studies, partial sequences of AR1 showed a high degree of similarity to the corresponding sequences of T4 phage (17, 37, 89, 90). In this study, we determined the complete genome sequence of AR1 (167,435 bp). The genome size of AR1 is highly similar to those of many other T4-like phages, including RB49, RB14, RB32, RB69, RB51, T4, JS98, JS10, 44RR, and RB43 (Table 1), but smaller than some large T4-like phages, for example, S-PM2, Aeh1, 65, KVP40, nt-1, and P-SSM2. AR1 is the largest known completely sequenced phage that infects the E. coli O157:H7 strain (see Table S1 in the supplemental material). The G+C content of the AR1 genome is 35.3%, which is within the G+C content range for T4-like phages (35.0% to 43.0%) but quite different from the G+C content (50.0%) of the E. coli O157:H7 host (Table 1). The G+C content of AR1 is smaller than those of the pseudot-even phages (average of 42.0%) and other sequenced E. coli O157:H7 phages (average of 48.3%) (see Table S1). This difference is in disagreement with previous studies which suggested that the G+C contents of phages and their hosts, such as the mycobacteriophages (60), Staphylococcus aureus phages (30), and E. coli phages (3) and their hosts, have similar values. However, the G+C content of lytic AR1 and T4-like phages might not be in agreement with the results from previous reports. To date, AR1 is the first completely sequenced and annotated T4-like phage of the E. coli O157:H7 strain.

Annotation of AR1 phage genes. Gene prediction (GeneMark.hmm and Zeevur_V) analysis revealed a total of 281 putative ORFs (encoding, on average, 189 amino acids [aa]) in the AR1 genome, which has 95% of its coding sequence carrying approximately 1.7 genes per kbp (Table 1). This property of the AR1 genome, i.e., highly packed genes, is consistent with previous findings for other phages, including the T4-like phages (21, 45, 60, 73). Upon Blastp and IPRscan analyses, 126 proteins (45%) could be annotated with known functions, 145 were conserved hypothetical proteins, including 36 proteins (13%) having protein domain information, and 10 (3%) were hypothetical proteins unique to the AR1 phage (Fig. 1A; see Table S5 in the supplemental material). Of the 281 putative ORFs, 271 ORFs (97%) had high protein sequence similarity to those of other T4-like phages. AR1 showed no protein sequence similarity to genes of other E. coli O157:H7 phages but had high protein sequence similarity to genes of T4-like phages (see Table S4). Further analysis of the AR1 proteins by Blastp searches against the proteins in the VOG database indicated that 126 ORFs (45%) could be classified into six VOG functional categories. Most of the named functional proteins, which are highly conserved among
AR1 shares a conserved T4-like core genome. It was reported previously that several gene sequences of AR1 are highly similar to those of T4 phage (17, 37, 89). To reveal the conservation among AR1 and T4-like phages, we compared the AR1 genome and seven closely related T4-like phage genomes. Sequence comparison indicated that AR1 had high nucleotide sequence similarity to the RB32, T4, RB69, and JS98 phages but little resemblance to the RB49, 44RR, and RB43 phages (Fig. 2). Nevertheless, AR1 has a very different genome sequence from those of other T4-like phages that can infect phylogenetically more distant bacteria, such as *Aeromonas*, *Acinetobacter*, *Vibrio*, and *Cyanobacteria* (data not shown).

It has been described that a conserved core genome comprising approximately 90 genes exists in the T4-like phages (8, 41, 44, 55, 62, 78). To elucidate the protein conservation among AR1 and other T4-like phages, a clustering analysis was performed using T4 phage proteins as the reference to identify the genes in various VOG functional clusters common to AR1 and the other sequenced T4-like phages. As shown in Fig. 3, AR1 had a conservation profile more similar to those of the T4, JS98, RB69, and RB32 phages, whereas the 44RR, RB43, and RB49 phages were more similar to each other than to AR1, suggesting that these phages can be divided into two groups. Despite the disparity of the two groups, there were still significant similarities between the T4-like phages. For instance, the phage RB43 shared 79.7% VOG similarity with T4 (Fig. 3). These conserved VOG clusters were also preserved in phages with different host ranges; for example, the *Aeromonas* phage 44RR that specifically infects *Aeromonas salmonicida* had 87.5% VOG similarity to the T4 phage that infects *E. coli* strains. This high VOG similarity suggests that T4-like phages apparently share a well-organized core genome containing the essential genes, which might have helped to keep T4-like phage genomes stable throughout the course of evolution.

Small differences, however, do exist in most of the VOG categories, except for the structural proteins (Fig. 3). For example, the genome modification enzyme dCMP hydroxymethylase (gp42) of the DNA replication gene category was found only in the T-even phages, while the DNA adenine methyltransferase (Dam) and some of the endonucleases were ubiquitously present in the pseudo-T-even phages. Despite these T4-like phages using similar DNA replication, repair system, nucleotide metabolism, and cellular machineries to propagate their genomes, they have different modification systems to help them adapt to different hosts. Another interesting finding is the presence of different transcription regulation-related proteins, including Alc (host transcription shutoff), MotA (activator of middle-mode transcription), MotB (modifier of transcription), ModB (ADP-ribosylase), and Srd (antisigma factor), that frequently exist in T-even phages. This implies that these proteins may use different factors to control or regulate viral gene expression.

Genome features of AR1 differ from those of other T4-like phages. Comparative analysis of AR1 and other T4-like phages indicated that there are several variations between the phage-encoded tRNAs, internal proteins (IPs), and mobile elements. First, upon the examination of tRNA genes in AR1, 10 tRNA encoded tRNAs, internal proteins (IPs), and mobile elements. First, upon the examination of tRNA genes in AR1, 10 tRNA genes were found clustered around the *segD* gene (Table 2). As expected, tRNA analysis of AR1 and other T4-like phages indicated that they have different numbers and types of tRNAs. This is in agreement with a previous study of tRNAs in T4-like phages (8). Further analysis of the codon usage frequencies in AR1 phage and *E. coli* O157:H7 suggested that at least five AR1-encoded tRNAs, the Leu (TAA), Met (ATG), Arg (AGA), Asn (CAA), and Tyr (TAC) tRNAs, may be related to decoding of the relatively more frequent codons in some of the AR1 genes (Table 2).
FIG. 3. Protein similarities of various VOG functional categories between AR1 and other T4-like phages. The abbreviations at the top indicate the six VOG functional categories as described in the legend to Fig. 1. Each horizontal gray line represents the conservation of a protein among the phages. The percentages at the bottom indicate the degrees of VOG similarity between T4 and the indicated phages. Phages could be classified into two groups, group I (AR1, RB69, RB32, JS98, and T4) and group II (RB49, 44RR, and RB43), based on VOG clustering. The genes indicated at the right (arrows) represent diverse patterns of specific VOG clusters between different phage groups. The category of unknown proteins had five diverse patterns between groups I and II.

To understand which AR1 genes could be enriched by AR1-carried tRNAs, we therefore analyzed the codon usage of 10 AR1 tRNAs. Figure S2 in the supplemental material shows the frequencies of the AR1 codons of the 10 phage-encoded tRNAs, with an average of 16.1%. Upon analysis of the top 10 and bottom 10 proteins containing the largest numbers of codons of these 10 AR1-encoded tRNAs by using the hypergeometric test, five structural proteins (50%; P < 0.001) of the VOG virion protein category were found in the top 10 protein list (see Table S3). This phenomenon did not exist in the bottom 10 protein list, among which most proteins (55%) were hypothetical proteins with no known function. Structural proteins of the AR1 phage are required components for the phage life cycle but are not necessities for the host. This may explain why AR1 uses these 10 AR1-encoded tRNAs to maintain the important functional module of structural proteins.

Second, most of the identified IP sequences from T4-like phages, which have been shown to play important roles in the early stage of infection to resist the host defense system and may subvert the specific host biosynthesis of regulatory or modifying functions (8, 9, 82), have a conserved consensus capsid targeting sequence (CTS) in the N-terminal region (8, 50, 67). The two predicted AR1 IPs, AR1_126 and AR1_146, had relatively low protein sequence similarities to the IPs of T4-like phages but relatively high similarities to the IP4 and IP5 proteins of RB14, RB32, and RB51 phages. The CTS pattern of AR1 IPs was identical to that of T-even IPs but different from those of the Aeromonas phages (44RR, 25, and 31). This difference may explain why AR1 uses these 10 AR1-encoded tRNAs to maintain the important functional module of structural proteins.

Another significant difference between the AR1 and T4 genomes is in the mobile elements in homing endonucleases and homologs of the T4 classification categories. Of the total of 15 T4 phage genes classified in this category, including seven seg, five mob, and three intron nuclease (I-TevI to -III) genes (75), only the SegE (site-specific endonuclease) homolog was found in AR1 phage (see Fig. S4 in the supplemental material). These genes of mobile elements do not commonly exist in the T4-like phages, and the differences in mobile elements between AR1 and other T4-like phages. All IP sequences were predicted using the same annotation pipeline as this study. A total of 22 putative IPs were identified in AR1 T4-like phage genomes. These IPs have a consensus N-terminal CTS motif, MKT(Y/F)(K/Q)EF (I/L)XE*. Asterisks denote the critical gp21 putative protease cutting sites in IPs. The IP-like protein sequences of the AR1 phage are similar to the IP sequences of T-even phages (RB14, RB32, and RB51) but different from those of the Aeromonas phages (44RR, 25, and 31). %A + K and pl indicate the percentage of alanine and lysine residues and the isoelectric point for each IP, respectively. The X in the consensus sequence represents any amino acid.

![Graphical representation of protein similarities](image)

**TABLE 2. tRNA genes and codon usage of AR1 phage**

<table>
<thead>
<tr>
<th>Locus tag</th>
<th>tRNA (anticodon)</th>
<th>Location (bp)</th>
<th>Length (bp)</th>
<th>GC%</th>
<th>Codon usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR1t001a</td>
<td>Arg (CTT)</td>
<td>68,329–68,404</td>
<td>76</td>
<td>55.3</td>
<td>+</td>
</tr>
<tr>
<td>AR1t002</td>
<td>His (GTG)</td>
<td>68,409–68,484</td>
<td>76</td>
<td>52.6</td>
<td>–</td>
</tr>
<tr>
<td>AR1t003</td>
<td>Asn (GTT)</td>
<td>68,598–68,672</td>
<td>75</td>
<td>50.7</td>
<td>+</td>
</tr>
<tr>
<td>AR1t004</td>
<td>Tyr (GTA)</td>
<td>68,677–68,763</td>
<td>87</td>
<td>55.2</td>
<td>+</td>
</tr>
<tr>
<td>AR1t005a</td>
<td>Met (CAT)</td>
<td>68,777–68,851</td>
<td>75</td>
<td>56.0</td>
<td>+</td>
</tr>
<tr>
<td>AR1t006b</td>
<td>Thr (TGT)</td>
<td>68,853–68,928</td>
<td>76</td>
<td>50.0</td>
<td>–</td>
</tr>
<tr>
<td>AR1t007</td>
<td>Ser (TGA)</td>
<td>68,934–69,023</td>
<td>90</td>
<td>57.8</td>
<td>–</td>
</tr>
<tr>
<td>AR1t008a</td>
<td>Gln (TTG)</td>
<td>69,685–69,758</td>
<td>74</td>
<td>41.9</td>
<td>–</td>
</tr>
<tr>
<td>AR1t009a</td>
<td>Leu (TAA)</td>
<td>69,764–69,850</td>
<td>87</td>
<td>48.3</td>
<td>+</td>
</tr>
<tr>
<td>AR1t010a</td>
<td>Gln (TTG)</td>
<td>69,851–69,924</td>
<td>74</td>
<td>45.9</td>
<td>+</td>
</tr>
</tbody>
</table>

a Three tRNA scan tools, namely, tRNAscan-SE, ARAGORN, and Rfam, were used for prediction. All of the predicted tRNA genes are located in the minus strand of the AR1 genome.

b Identified in both the AR1 and T4 (45) genomes.

c Codon usage was analyzed with the CUSP and CAI programs of EMBSS version 6.2.0 (68).

![Graphical representation of characteristic of predicted IPs of AR1 and other T4-like phages](image)
AR1 and T4 phage genomes is consistent with the results for another T4-like phage, RB49 (8). Although a few group I-like introns have been found in other phages, such as T4 phage, *Bacillus* phages SPO1 and Bastille (31), *S. aureus* phage Twort (32), *Lactobacillus* phage LL-H (43), and *Lactococcus* phage r1t (84), none of the three T4 homing endonucleases or the group I-like introns could be found in AR1 (see Fig. S4).

**Computational analysis of AR1 virion proteins.** The AR1 structural proteins were predicted by comparing the predicted proteins with the 43 possible T4 structural proteins identified by LC-MS/MS. (B) Purified phage particles of AR1, T4, and RB69 phages were resolved in 10% SDS-PAGE gels and visualized by silver staining. Lanes 2, 3, and 4, AR1, T4, and RB69 virion proteins, respectively; lane 1, protein standard markers. The sizes (in kDa) of the AR1 structural proteins are shown on the left, and the band at 52 kDa is a major head protein of the AR1, T4, and RB69 phages. (C) Mass spectrometry-identified virion proteins.
in a previous study (45). Using cutoffs of 30% sequence identity, 70% sequence coverage, and an E value of \( < 10^{-4} \), 38 AR1 proteins shared sequence similarity with T4 essential structural components (Fig. 5A). There were five T4 virion proteins, which included the head protein (gp67), receptor recognition protein (gp37), and three IPs (Ipf, -II, and -III), having relatively low sequence coverage and/or similarity with their AR1 counterparts (Fig. 5A; see Fig. S4 in the supplemental material). The AR1 head protein (gp67) had part of its sequence that was highly similar with the T4 homolog (96% identity) but had a low level of coverage (63%). Another notable difference in virion proteins was in the receptor recognition protein gp37, which had 32.2% identity and 67.7% sequence coverage between the AR1 and T4 phages.

**Proteome analysis of AR1 phage particles.** The structural proteins of purified AR1 and the related T4 and RB69 phage particles were examined by SDS-PAGE. Approximately 24 visible protein bands were observed in the AR1 lane (Fig. 5B, lane 2). Unlike T4 and RB69, which had similar protein profiles, several AR1 proteins seemed to have different molecular weights (especially those larger than the major capsid protein) and did not match those in the T4 or RB69 phage (Fig. 5B). On the other hand, upon liquid chromatography (LC)-MS/MS analysis of trypsin-digested AR1 phage proteins, 18 structural proteins (ProteinProphet probability of \( \geq 0.9 \)) were identified with protein-specific unique peptides (PeptideProphet probability of \( \geq 0.9 \)) (Fig. 5C). The top 5 identified viral structural proteins were gp23, IP4, gp18, gp34, and soc. All of the identified structural proteins included components of the head, tail, tail fibers, and baseplate (Fig. 5A, bold and underlined proteins, and C), in agreement with their structural function predicted in our genome annotations and also in agreement with previous publications (8, 35, 37, 45, 89, 90). However, our LC-MS/MS analysis also identified several proteins, including two, RNA polymerase ADP-riboseylase (alt) and holin (\( \alpha \)), which had not been annotated as structural proteins. The Alt gene product, which plays an important role in shutting off host transcription, is a virion component incorporated into the phage head (69). The holin protein (involved in host membrane lysis) may be expressed at a high level in the late stage of infection to damage the host membrane (88). It is possible that these proteins may have emerged and been preserved in our proteome data due to copurification with the phage particles.

A conserved hypothetical protein of unknown function (AR1_137) which is highly similar to RB32ORF133c (YP_803075) of RB32 phage was also detected by LC-MS/MS. These two proteins have a CTS pattern (MKSYAQFLXE*) similar to the T-even CTS and a gp21 protease recognition site in their N-terminal regions. The sequence properties of AR1_137 are also consistent with those of the other IPs (%A\( + K \), 31%; and pI, 10.1). Most of the IP genes are located between the genes vs.8 and vs.e and/or trmA.4 and 57B in T4 and RB69 phages (8). The genomic loci of AR1_137 and RB32ORF133c are not present in the common location (between vs.8 and vs.e or trmA.4 and 57B) for IPs in T-even phages but are very close to the trmA.4 gene. Based on the results of sequence and proteome analyses, we considered that they may belong to IPs incorporated inside the phage head.

**Sequence diversity between putative receptor recognition domains of AR1 gp37 and gp37 of related T4-like phages.** Both AR1 and PP01 phages, which are known to possess gp37, have a host specificity different from that of the other known T4-like phages with the same protein. Previous studies demonstrated that gp37 in T4 phage is responsible for binding to the host receptor OmpC on *E. coli* K-12 (15, 48, 51) and that deletion of OmpC in the O157:H7 strain would abolish AR1 infection (89). If AR1 adsorption is also dependent on the binding of gp37 protein to OmpC, then specificity could result from protein sequence variations in either gp37, OmpC, or both. Host infection experiments performed by us and previous studies (49, 89, 92) indicated that phages AR1 and PP01 could infect O157:H7 but not K-12 and that phages RB69, JS98, and T4 could infect K-12 but not O157:H7 (see Fig. S6A in the supplemental material). Therefore, the gp37 and OmpC proteins from the O157:H7 strain-specific (AR1 and PP01) and K-12 strain-specific (RB69, JS98, and T4) phages may be divided into two groups by multiple sequence alignment analysis.

AR1 gp37 has high sequence similarity to the gp37 proteins of PP01, RB69, and JS98 phages but lower similarity to that of T4 phage (see Fig. S6B in the supplemental material). Several highly conserved regions were found in the five T4-like phages (Fig. 6A, peptide illustration). The majority of the conserved blocks were found in the central regions of gp37 homology. Another conserved block was found in the first 100 amino acids of the N-terminal region, a result consistent with a previous study of gp37 proteins of T-even phages (80). Previous studies of T4 gp37 suggested that these conserved regions do not seem to be involved in the binding of gp37 to OmpC (22, 48). However, the sequence near the last \(~130\) amino acids in the C terminus of T4 gp37 contributes most to the host specificity of T4 infection (22, 48, 57, 80, 90). To identify sequence variations between the receptor recognition areas of gp37 C terminus, we further analyzed this region of the five gp37 proteins. Multiple sequence alignment indicated that eight residues in the gp37 C terminus are identical in both AR1 and PP01 but different from those in phages RB69, JS98, and T4. There are seven varied residues centralized and located within a segment (aa 939 to 972) of the AR1 gp37 C-terminal region (Fig. 6A, asterisks). Interestingly, the varied residues are located within the T4 gp37 receptor recognition sequence (22, 48). We speculated that the eight varied residues may be associated with binding to the O157:H7 OmpC receptor.

On the other hand, the sequence variations between the *E. coli* K-12 and O157:H7 OmpC receptor proteins may also play a role in host specificity. Comparison of the two O157:H7 phage strains (AR1 and PP01) did not seem to be conclusive. Multiple sequence alignment suggested that several sequence variations could be found between the OmpC proteins of O157:H7 and K-12, including substitutions of 14 amino acids and two insertion/deletion areas (Fig. 6B). Previous studies (46, 83) indicated that eight point mutations of OmpC can make K-12 resistant to T4-like phage infection (Fig. 6B, asterisks). The major varied region (aa 176 to 184 of K-12) of the OmpC proteins of O157:H7 and K-12 was much closer to the potential gp37 binding sites. Thus, it is speculated that sequence variations of both the gp37 receptor recognition domain and OmpC binding sites may contribute to AR1 phage specificity for the *E. coli* O157:H7 strain.
FIG. 6. Comparison of gp37 protein sequences of T4-like phages AR1, PP01, RB69, JS98, and T4. (A) The gp37 conserved regions of AR1 and five other T4-like phages are indicated by gray boxes in the peptide illustration. The coordinates (in aa) next to the boxes indicate the positions of the gp37 conserved regions in AR1. The C-terminal regions (corresponding to the receptor recognition area of T4 phage) of the gp37 proteins of five T4-like phages were used for multiple sequence alignment. The asterisks show those amino acid residues of AR1 that are identical to the corresponding residues of PP01 but different from those of RB69, T4, or JS98 phage. (B) Sequence comparison of OmpC host receptor proteins between E. coli K-12 MG1655, O157:H7 1266, and O157:H7 ATCC 43888, which are the hosts of T4, AR1, and PP01 phages, respectively. The asterisks indicate the specific substitutions and insertions/deletions of eight residues that resulted in the loss of binding activity of the gp37 protein to the OmpC receptor (46, 83). The sequence analysis was performed with CLC Sequence Viewer 5.1.2 (CLC bio).
DISCUSSION

The complete genome of AR1 phage, which can efficiently infect clinical isolates of *E. coli* O157:H7, was sequenced completely in this study. Unexpectedly, AR1 has no shared genes with any other currently sequenced *E. coli* O157:H7 phages. The genomic features of the AR1 phage are also extremely different from those of other *E. coli* O157:H7 phages in terms of genome size, G+C content, gene number, whole-genome sequence, and proteome (see Fig. S1 and Tables S1 and S4 in the supplemental material). Comparative genomic analysis indicated that AR1 has a very similar genome organization and core genome to those of seven sequenced T4-like phages, especially the T4, RB32, RB69, and JS98 phages (see Fig. S5). It is clearly displayed in Fig. 3 and 7 that AR1 has a conserved core genome with seven other T4-like genomes. Structural protein and DNA replication-related genes are clustered in the first third and the late regions of the T4-like genomes (see Fig. S5), respectively. This finding is consistent with previous observations of T4-like phages (5, 8, 12, 55, 62, 92). In addition, there are conserved regions encoding proteins with unknown function between these two highly conserved modules in these phage genomes (see Fig. S5). These conservations suggest that T4-like phages use similar mechanisms to control propagation in their hosts.

Several diverse features between AR1 and the other T4-like phages, including the phage-encoded tRNA*s, I*Ps, and mobile elements, were also observed in the genome comparison results. An interesting difference was seen in the types and numbers of phage-encoded tRNA*s, which may be related to their relatively high codon frequency in some phage-encoded proteins. Previous studies indicated that T4-encoded tRNA*s are related to codons that are highly used in T4 genes but rarely used in the host and that the phage tRNA*s can enhance the lowly expressed T4 late-stage protein genes through optimal codon usage in translation (29, 45). Similarly, analysis of the AR1-encoded tRNA*s indicated that half of them may be related to the optimal codon usage of certain AR1 proteins (Table 2). Another study of T4 phage observed that deletion of the T4-encoded tRNAs would reduce both the phage burst size and T4 protein synthetic rate (86). Our further analysis of codon usage showed that most of the AR1 structural proteins may have optimal codons related to the 10 AR1-encoded tRNA*s and that their usage percentages of these tRNA*s are dramatically higher than the average (16.1%) for the entire AR1 proteome. It is speculated that phages must carry and
keep the various numbers and types of phage-carried tRNA genes in their genomes through evolutionary processes to optimize the expression of the late genes which encode the structural proteins.

Protein sequence analyses of the IPs were in accordance with a previous report that enormous hypervariable sequences of the IP genes exist in T4-like phages (8). Our results showed that AR1 has two IP sequences that contain a conserved N-terminal CTS, MKT(Y/F)(K/Q)EF(I/L)XE*, similar to those of other IPs encoded by different T4-like phages. This CTS pattern is more similar to those of T4, RB14, RB32, RB51, RB69, JS10, and JS98 phages (Fig. 4) and more distant from those of the T4-like group of Aeromonas phases (8). The differences in IPs may be related to resistance to different host defense systems, such as the type IA restriction-modification (RM) system in E. coli K-12 and K1 (and also in S. Typhimurium LT2) and the type IB RM in pathogenic E. coli O157:H7 and CFT073 strains (8, 9). Different IPs may be specific for sabotaging various types of host RM defense systems to avoid being attacked by host restriction enzymes after phage genome injection.

The distal tail fibers of the T4-like phages contain the protein products of genes 36, 37, and 38, which are known to be the major determinants of specificity for adsorption to bacterial hosts (22, 25, 80). Previous studies demonstrated that the T4 phage receptor recognition site could be defined approximately 130 residues from the T4 gp37 C terminus (22, 48, 80). Analysis of AR1 gp37 identified eight varied residues in the putative receptor recognition region (Fig. 6A, bottom panel) near the receptor recognition sites of T4 gp37 (22). These eight varied amino acids in the hypervariable C-terminal region of AR1 gp37 may be related to recognition of O157:H7 OmpC. We also found slight sequence differences between O157:H7 and K-12 OmpC receptors (Fig. 6B) in locations which are associated with gp37 binding activity in several T4-like phages (such as T4, Tula, and TUb). These findings may relate to AR1 specificity for the E. coli O157:H7 strain. Additional analysis of the T4-like gp37 proteins found several leucine and isoleucine zipper motifs, L-6X-L-6X-L and L-6X-I-L-6X-I-L-6X-I, at the AR1 gp37 C terminus (see Fig. S6C in the supplemental material). These zipper motifs were also found in the other T-even phages, including T2, K3, T6, Ox2, M1, and Ac3 (80).

A previous study indicated that a single point mutation (A1018T) in the T4 gp37 zipper motif region led to a loss of binding activity to the OmpC receptor (22). The functions of leucine and isoleucine zipper motifs may be associated with the trimerization of gp37 proteins to allow them to form the proper structure to interact with the cellular receptor OmpC. We speculated that the A1018T substitution may affect the gp37 protein conformation and/or assembly but may not be necessary for binding to the OmpC receptor.

There are currently more than 40 T4-like genomes available in GenBank, the GGC T4 genome database (http://phage.ggc .edu), and the literature (10, 64, 79, 92). To address the relationships between AR1 and the newly published T4-like phage genomes, we further analyzed the similarities of genome sequences, proteomes, and tRNA genes between AR1 and the other T4-like phages. The tRNA analysis showed that most AR1 tRNA sequences have similar tRNA counterparts in RB18, RB14, RB26, RB51, RB32, T2, RB15, T4, T4T, and T6 (see Fig. S3 and Table S2B in the supplemental material). There is no sequence similarity between tRNAs of AR1 and RB69 and low similarity with those of JS10, CC31, and JS98. By increasing the Blastp identity cutoff value, RB69, JS98, and JS10 can be discriminated from the T-even subtype (see Fig. S7). However, both Dotter analysis (Fig. 2, thick box) and VOG functional clustering analysis (Fig. 3) showed that AR1 is a T4-related phage similar to RB14, RB32, RB69, RB51, T4, JS98, and JS10 phages. The phylogenetic tree analysis also showed that RB51, T4, RB14, RB32, RB69, and JS98 can be classified into the same T4 subtype (unpublished data). This result is similar to those of a study published by Lavigne et al. in 2009 (34). Therefore, based on the results of previous studies (34, 63) and our analyses, AR1 has a much greater similarity to T4 and other T-even phages than to RB69, JS98, and JS10 phages. The RB69, JS98, and JS10 were shown to be more distant from the AR1 and T-even phages, similar to a study performed by Petrov et al. (63).

E. coli O157:H7 infection is a significant human health threat. There were 350 outbreaks representing 8,598 cases of E. coli O157:H7 infection between 1982 and 2002 in the United States (65). Establishing methods to control E. coli O157:H7 contamination has been an important safety issue. Several groups reported the use of bacteriophage cocktails as a means of biological control to remove contamination of pathogenic E. coli O157:H7 (1, 28, 56). These studies demonstrated that naturally occurring phages have extremely high efficiencies for eliminating E. coli O157:H7 from contaminated foods or container surfaces. AR1 can lyse 68 STEC strains, 38 ECOR strains (17), and 28 clinical strains of E. coli O157:H7 (89). Unlike the other E. coli O157:H7 phages, which can produce Shiga-like toxin during the lysogenic stage, no Shiga-like or other toxins were found to be encoded by the AR1 genome. Therefore, AR1 phage has the potential to be used as a bacterial food additive to prevent contamination with E. coli O157:H7 in the food industry. In addition, several reports have indicated that some antibiotics may increase Shiga-like toxin production, resulting in increased side effects for patients (39, 61). Therefore, AR1 could also be used as an alternative therapeutic agent for treating E. coli O157:H7 infection. Finally, an understanding of AR1 and T4-like phage genomes can help to bring insights into biological control of pathogen contamination and genome engineering for practical use in the future.

ACKNOWLEDGMENTS

We thank Yi-Feng Chang, Ying-Hsueh Huang, and Chi Yang for their help with bioinformatics analysis and Hung-Yu Shu for DNA sequencing in the Genome Research Center at National Yang Ming University.

This work was supported by grants NSC98-2511-S-010-004-MY2 and NSC99-3112-B-010-018 from the National Research Program for Genomic Medicine and the National Science Council, Taiwan, and by a grant from the Ministry of Education, Aim for the Top University Plan, to the National Yang Ming University.

REFERENCES


