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,⁵ and Chuan-Hsiung Chang^{2,3*}

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 YYZ-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 acces-

sion 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 Ronner 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 Shiga toxin-producing *E. coli* (STEC) and *E. coli* collection of reference (ECOR) strains but not the *E. coli* K-12 or B^E strains. In addition, AR1 can also lyse other *Enterobacteriaceae* bacteria, such as *Shigella dysenteriae* and *Salmonella enterica* subsp. *enterica* 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-

* Corresponding author. Mailing address: Institute of Biomedical Informatics, National Yang-Ming University, 155, Sec. 2, Linong St., Taipei 11221, Taiwan. Phone: 886-2-28267316. Fax: 886-2-28206754. E-mail: cchang@ym.edu.tw.

[†] Supplemental material for this article may be found at <http://jvi.asm.org/>.

[‡] Published ahead of print on 20 April 2011.

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], ϵ 11/2 [56], PP01 [49, 56], and ECML-134 [1]) have also been discovered to efficiently infect and lyse the *E. coli* O157:H7 strain. 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^E (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^E (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* DH5 α . 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) were obtained 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 pipeline integrates the gene prediction results from GeneMark.hmm 2.0 (version 2.6) (39) and Zcurve_V (version 1.0) (19), with a minimal size of putative ORFs of 90 nucleotides, equivalent to the coding region for a minimal polypeptide 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 query/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, RB32, 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 analysis. The viral structural proteins were treated with 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) sequences 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 ($P \geq 0.9$) 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

Phage ^a	GenBank accession no.	Genome size (bp)	No. of predicted ORFs	GC%	Gene density ^b	% Coding sequence	Avg gene product size (aa) ^c
RB49	NC_005066	164,018	279	40.0	1.7	94	184
RB14	NC_012638	165,429	274	35.0	1.7	94	189
RB32	NC_008515	165,890	270	35.0	1.6	94	193
AR1	AP011113	167,435	281	35.3	1.7	95	189
RB69	NC_004928	167,560	273	37.0	1.6	93	190
RB51	NC_012635	168,394	273	35.0	1.6	94	193
T4	NC_000866	168,903	278	35.0	1.6	93	188
JS98	NC_010105	170,523	266	39.0	1.6	91	194
JS10	NC_012741	171,451	265	39.0	1.5	92	198
44RR	NC_005135	173,591	252	43.0	1.5	92	211
RB43	NC_007023	180,500	292	43.0	1.6	93	192

^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^E 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.bioc.tulane.edu>), and GGC T4-like genome (<http://phage.ggc.edu>) websites.

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

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

GGC 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 phage genomes 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, <10⁻⁴).

Nucleotide sequence accession numbers. The nucleotide sequence of phage AR1 has been deposited in the DDBJ database under accession number AP011113. 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_004928 (55, 62); RB51, NC_012635 (Nolan et al., direct submission); T4, NC_000866 (45); JS98, NC_010105 (92); JS10, NC_012741 (10); 44RR, NC_005135 (55, 81); and RB43, NC_007023 (55, 62). The GenBank accession numbers for internal protein sequences are as follows: T4 IP1, NP_049749; T4 IP2, NP_049734; T4 IP3, NP_049735; RB14 IP1, YP_002854476; RB14 IP4, YP_002854462; RB14 IP5, YP_002854477; RB32 IP4, YP_803065; RB32 IP5, YP_803083; RB32 IP7, YP_803081; RB32 IP9, YP_803082; RB51 IP4, YP_002854082; RB51 IP5, YP_002854099; RB69 IP1, NP_861845; RB69 IP5, NP_861817; RB69 IP7, NP_861846; RB69 IP10, NP_861814; JS10 IP3, YP_002922462; JS10 IP12, YP_002922461; JS98 IP3, YP_001595244; and JS98 IP14, YP_001595243. The GenBank accession numbers for gp37 protein sequences are as follows: AR1, BAI83266; PP01, AAK30164; T4, NP_049863; RB69, NP_861944; and JS98, YP_001595372. 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 pseudo-T-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 contents 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 Zcurve_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

FIG. 1. Summary of the 281 AR1 genes annotated by Blastp and InterProScan analyses. (A) Pie chart showing the relative abundances of genes with and without named functions, including 126 KFP (known function protein), 109 CHP (conserved hypothetical protein), 36 CHP-D (CHP with protein domain), and 10 HP (hypothetical protein) genes. (B) Relative abundances of AR1 proteins in the six VOG functional categories. T, RNA replication, transcription, and modification; A, auxiliary proteins; L, DNA replication and repair and nucleotide metabolism; R, regulation of cellular metabolism; S, structural proteins; U, unknown proteins. NAF, ORFs that cannot be assigned to any functional categories.

T4-like phages, belonged to the structural proteins (36 ORFs) and DNA replication proteins (34 ORFs) (Fig. 1B).

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

FIG. 2. Genome sequence similarities between AR1 and other T4-like phages. The Dotter program (76) was used with a sliding window of 25 bp to align the phage genome sequences according to their genome sizes. Genome sequences showing extremely high similarity are boxed (thick line) and belong to T4-type phages (34, 63). The phages from other subtypes (RB49, 44RR, and RB43) have relatively low sequence similarities to AR1.

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 (anti-sigma 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 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

TABLE 2. tRNA genes and codon usage of AR1 phage

Locus tag	tRNA (anticodon) ^a	Location (bp)	Length (bp)	GC%	Codon usage ^c
AR1t001 ^b	Arg (TCT)	68,329–68,404	76	55.3	+
AR1t002	His (GTG)	68,409–68,484	76	52.6	–
AR1t003	Asn (GTT)	68,598–68,672	75	50.7	+
AR1t004	Tyr (GTA)	68,677–68,763	87	55.2	+
AR1t005 ^b	Met (CAT)	68,777–68,851	75	56.0	+
AR1t006 ^b	Thr (TGT)	68,853–68,928	76	50.0	–
AR1t007 ^b	Ser (TGA)	68,934–69,023	90	57.8	–
AR1t008 ^b	Gly (TCC)	69,685–69,758	74	41.9	–
AR1t009 ^b	Leu (TAA)	69,764–69,850	87	48.3	+
AR1t010 ^b	Gln (TTG)	69,851–69,924	74	45.9	–

^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 EMBOSS, version 6.2.0 (68). +, AR1-carried tRNA genes using different frequencies of optimal codons with respect to the host (*E. coli* O157:H7); –, all other genes.

Phage	ORF	N-terminal CTS	aa	%A+K	pI
AR1	IP4	MKT Y Q E F I T E *AA	157	20	9.4
	146	MKT F K E F A T K T I T E *SS	85	28	10.1
T4	IP1	MKT F K E F T S T T I P V S T I T E*AT	95	23	9.7
	IP2	MKT Y Q E F I A E *AR	100	31	10.0
	IP3	MKT Y Q E F I A E *AS	193	27	10.2
RB14	IP1	MKT F K E F T S T T I P V S T I T E*AT	95	23	9.7
	IP4	MKT Y Q E F I T E *AA	157	20	9.4
	IP5	MKT F K E F A T K T I T E *SS	85	28	10.1
RB32	IP4	MKT Y Q E F I T E *AA	157	20	9.4
	IP5	MKT F K E F A T K T I T E *SS	85	28	10.1
	IP7	MKT F K E F I N E *AA	102	26	10.3
	IP9	MKT Y Q E F I S E *AA	90	19	7.5
RB51	IP4	MKT Y Q E F I T E *AA	157	20	9.4
	IP5	MKT F K E F A T E T I T E *SS	85	27	9.2
RB69	IP1	ME P K M K T Y K E F I S P V A N V S T L T E *AT	99	23	9.5
	IP5	MKT Y M N F Y Q N L S E N E*AT	93	22	8.5
	IP7	MKT Y Q E F I A E *AT	108	19	10.0
	IP10	MKT Y K E F I A E *SA	90	13	6.8
JS10	IP3	MKT Y K E F I T E *AT	198	27	10.2
	112	MKT Y N E F I A E *AV	132	20	6.9
JS98	IP3	MKT Y K E F I T E *AT	198	28	10.2
	114	MKT Y N E F I A E *AV	132	20	6.9
Consensus sequence		MKT Y K E F I X E * FQ L			

FIG. 4. Characteristics of predicted IPs of AR1 and other T4-like phages. All IP sequences were predicted using the same annotation pipeline in this study. A total of 22 putative IPs were identified in the T4-like phage genomes. These IPs have a consensus N-terminal CTS pattern, 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 pI 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.

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 the 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 in *Aeromonas* phages (Fig. 4). The differences in the AR1 IPs may be related to resistance to the *E. coli* O157:H7 host defense system or to adaptation.

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 difference in mobile elements between

FIG. 5. Proteomic and computational analyses of AR1 virion proteins. (A) Computational alignment and mapping of structural components of the AR1 phage to T4 structural proteins based on protein sequence similarity (identity of >30%, coverage of >70%, and E value of <10⁻⁴). The AR1 virion comprises five major components: the head, neck, tail, baseplate, and tail fibers (17). The numbers inside the boxes are the AR1 ORFs matched to T4 structural proteins. Bold and underlined numbers denote the AR1 virion 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.

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

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 (IpI, -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 ≥ 0.9) were identified with protein-specific unique peptides (PeptideProphet probability of ≥ 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-ribosylase (*alt*) and holin (*t*), 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 (MKSQAFLXE*) 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 *tma.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 *tma.4* and *57B*) for IPs in T-even phages but are very close to the *tma.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 termini, 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. 7. Conserved functional modules of structural protein genes (A) and DNA replication-related genes (B) in the T4-like phages. The eight phages were arranged according to genome size. The annotated gene names for T4 phage were followed for the gene names (arrows) of RB49 phage. The Blastp program was first used for protein sequence comparisons with the following cutoffs: identity, >30%; coverage, >70%; and E value, $<10^{-4}$. Lines were then drawn to connect the homologous genes (arrows not drawn to scale) in these phages. Several nonconserved genes can be found in panel A (black arrows): (i) gene 67 of the AR1, RB32, and RB69 phages; (ii) gene 24 is duplicated in JS98 phage; and (iii) RB43 has a nonconserved region between genes 24 and 48. Other varied organizations are identified in panel B (black arrows): (i) gene 42, (ii) genes 46 and 47 of the RB43 phage, and (iii) the splicing intron genes *nrdB* and *td* of the T4 phage.

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 tRNAs, IPs, and mobile elements, were also observed in the genome comparison results. An interesting difference was seen in the types and numbers of phage-encoded tRNAs, which may be related to their relatively high codon frequency in some phage-encoded proteins. Previous studies indicated that T4-encoded tRNAs are related to codons that are highly used in T4 genes but rarely used in the host and that the phage tRNAs can enhance the lowly expressed T4 late-stage protein genes through optimal codon usage in translation (29, 45). Similarly, analysis of the AR1-encoded tRNAs 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 tRNAs and that their usage percentages of these tRNAs 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* phages (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, TuIa, and TuIb). 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 I-6X-I-6X-I-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 bactericidal 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

1. Abuladze, T., et al. 2008. Bacteriophages reduce experimental contamination of hard surfaces, tomato, spinach, broccoli, and ground beef by *Escherichia coli* O157:H7. *Appl. Environ. Microbiol.* **74**:6230–6238.
2. Ackermann, H. W. 2007. 5500 phages examined in the electron microscope. *Arch. Virol.* **152**:227–243.
3. Ackermann, H. W., and M. S. DuBow. 1987. Viruses of prokaryotes. CRC Press, Boca Raton, FL.

4. Ackermann, H. W., and H. M. Krisch. 1997. A catalogue of T4-type bacteriophages. *Arch. Virol.* **142**:2329–2345.
5. Arbiol, C., A. M. Comeau, M. Kutateladze, R. Adamia, and H. M. Krisch. 2010. Mobile regulatory cassettes mediate modular shuffling in T4-type phage genomes. *Genome Biol. Evol.* **2**:140–152.
6. Belshaw, R., and A. Katzourakis. 2005. BlastAlign: a program that uses Blast to align problematic nucleotide sequences. *Bioinformatics* **21**:122–123.
7. Chibani-Chennoufi, S., C. Canchaya, A. Bruttin, and H. Brussow. 2004. Comparative genomics of the T4-like *Escherichia coli* phage JS98: implications for the evolution of T4 phages. *J. Bacteriol.* **186**:8276–8286.
8. Comeau, A. M., C. Bertrand, A. Letarov, F. Tetart, and H. M. Krisch. 2007. Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* **362**:384–396.
9. Comeau, A. M., and H. M. Krisch. 2005. War is peace—dispatches from the bacterial and phage killing fields. *Curr. Opin. Microbiol.* **8**:488–494.
10. Denou, E., et al. 2009. T4 phages against *Escherichia coli* diarrhea: potential and problems. *Virology* **388**:21–30.
11. Desplats, C., C. Dez, F. Tetart, H. Eleaume, and H. M. Krisch. 2002. Snapshot of the genome of the pseudo-T-even bacteriophage RB49. *J. Bacteriol.* **184**:2789–2804.
12. Desplats, C., and H. M. Krisch. 2003. The diversity and evolution of the T4-type bacteriophages. *Res. Microbiol.* **154**:259–267.
13. Elias, J. E., and S. P. Gygi. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**:207–214.
14. Fraser, C. M., and R. D. Fleischmann. 1997. Strategies for whole microbial genome sequencing and analysis. *Electrophoresis* **18**:1207–1216.
15. Furukawa, H., and S. Mizushima. 1982. Roles of cell surface components of *Escherichia coli* K-12 in bacteriophage T4 infection: interaction of tail core with phospholipids. *J. Bacteriol.* **150**:916–924.
16. Gan, R. R., et al. 2006. Proteome analysis of *Halobacterium* sp. NRC-1 facilitated by the biomodule analysis tool BMSorter. *Mol. Cell. Proteomics* **5**:987–997.
17. Goodridge, L., A. Gallaccio, and M. W. Griffiths. 2003. Morphological, host range, and genetic characterization of two coliphages. *Appl. Environ. Microbiol.* **69**:5364–5371.
18. Griffiths-Jones, S., et al. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**:D121–D124.
19. Guo, F. B., and C. T. Zhang. 2006. ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinformatics* **7**:9.
20. Gyles, C. L. 2007. Shiga toxin-producing *Escherichia coli*: an overview. *J. Anim. Sci.* **85**:E45–E62.
21. Hardies, S. C., J. A. Thomas, and P. Serwer. 2007. Comparative genomics of *Bacillus thuringiensis* phage 0305phi8-36: defining patterns of descent in a novel ancient phage lineage. *Virol. J.* **4**:97.
22. Hashemolhosseini, S., D. Montag, L. Kramer, and U. Henning. 1994. Determinants of receptor specificity of coliphages of the T4 family. A chaperone alters the host range. *J. Mol. Biol.* **241**:524–533.
23. Izumiya, H., et al. 1997. Molecular typing of enterohemorrhagic *Escherichia coli* O157:H7 isolates in Japan by using pulsed-field gel electrophoresis. *J. Clin. Microbiol.* **35**:1675–1680.
24. Karmali, M. A., V. Gannon, and J. M. Sargeant. 2010. Verocytotoxin-producing *Escherichia coli* (VTEC). *Vet. Microbiol.* **140**:360–370.
25. Kellenberger, E., et al. 1965. Functions and properties related to the tail fibers of bacteriophage T4. *Virology* **26**:419–440.
26. Keller, A., J. Eng, N. Zhang, X. J. Li, and R. Aebersold. 2005. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.* **1**:2005.0017.
27. Keller, A., A. I. Nesvizhskii, E. Kolker, and R. Aebersold. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74**:5383–5392.
28. Kudva, I. T., S. Jelacic, P. I. Tarr, P. Yoderian, and C. J. Hovde. 1999. Biocontrol of *Escherichia coli* O157 with O157-specific bacteriophages. *Appl. Environ. Microbiol.* **65**:3767–3773.
29. Kunisawa, T. 1992. Synonymous codon preferences in bacteriophage T4: a distinctive use of transfer RNAs from T4 and from its host *Escherichia coli*. *J. Theor. Biol.* **159**:287–298.
30. Kwan, T., J. Liu, M. DuBow, P. Gros, and J. Pelletier. 2005. The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proc. Natl. Acad. Sci. U. S. A.* **102**:5174–5179.
31. Landthaler, M., B. W. Shen, B. L. Stoddard, and D. A. Shub. 2006. I-Basi and I-Hmul: two phage intron-encoded endonucleases with homologous DNA recognition sequences but distinct DNA specificities. *J. Mol. Biol.* **358**:1137–1151.
32. Landthaler, M., and D. A. Shub. 1999. Unexpected abundance of self-splicing introns in the genome of bacteriophage Twort: introns in multiple genes, a single gene with three introns, and exon skipping by group I ribozymes. *Proc. Natl. Acad. Sci. U. S. A.* **96**:7005–7010.
33. Laslett, D., and B. Canback. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* **32**: 11–16.
34. Lavigne, R., et al. 2009. Classification of Myoviridae bacteriophages using protein sequence similarity. *BMC Microbiol.* **9**:224.
35. Leiman, P. G., S. Kanamaru, V. V. Mesyanzhinov, F. Arisaka, and M. G. Rossmann. 2003. Structure and morphogenesis of bacteriophage T4. *Cell. Mol. Life Sci.* **60**:2356–2370.
36. Liangke, S., Y. Yaxian, and L. Chengping. 2008. Construction of a stx2 deletion mutant of Shiga toxin 2 phage phiMin27 and its infectious properties. *Wei Sheng Wu Xue Bao* **48**:1227–1233.
37. Liao, C. P., and W. Syu, Jr. 2002. Analysis of the baseplate region of phage AR1 that specifically infects *Escherichia coli* O157:H7. *J. Microbiol. Immunol. Infect.* **35**:269–271.
38. Lowe, T. M., and S. R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**:955–964.
39. Lukashin, A. V., and M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26**:1107–1115.
40. Makino, K., et al. 1999. Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic *Escherichia coli* O157:H7 derived from the Sakai outbreak. *Genes Genet. Syst.* **74**:227–239.
41. Mann, N. H., et al. 2005. The genome of S-PM2, a “photosynthetic” T4-type bacteriophage that infects marine *Synechococcus* strains. *J. Bacteriol.* **187**: 3188–3200.
42. Margulies, M., et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376–380.
43. Mikkonen, M., and T. Alatossava. 1995. A group I intron in the terminase gene of *Lactobacillus delbrueckii* subsp. *lactis* phage LL-H. *Microbiology* **141**:2183–2190.
44. Miller, E. S., et al. 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J. Bacteriol.* **185**:5220–5233.
45. Miller, E. S., et al. 2003. Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **67**:86–156.
46. Misra, R., and S. A. Benson. 1988. Genetic identification of the pore domain of the OmpC porin of *Escherichia coli* K-12. *J. Bacteriol.* **170**:3611–3617.
47. Miyamoto, H., et al. 1999. Sequence analysis of Stx2-converting phage VT2-Sa shows a great divergence in early regulation and replication regions. *DNA Res.* **6**:235–240.
48. Montag, D., S. Hashemolhosseini, and U. Henning. 1990. Receptor-recognizing proteins of T-even type bacteriophages. The receptor-recognizing area of proteins 37 of phages T4 TuLa and TuB. *J. Mol. Biol.* **216**:327–334.
49. Morita, M., et al. 2002. Characterization of a virulent bacteriophage specific for *Escherichia coli* O157:H7 and analysis of its cellular receptor and two tail fiber genes. *FEMS Microbiol. Lett.* **211**:77–83.
50. Mullaney, J. M., and L. W. Black. 1996. Capsid targeting sequence targets foreign proteins into bacteriophage T4 and permits proteolytic processing. *J. Mol. Biol.* **261**:372–385.
51. Mutoh, N., H. Furukawa, and S. Mizushima. 1978. Role of lipopolysaccharide and outer membrane protein of *Escherichia coli* K-12 in the receptor activity for bacteriophage T4. *J. Bacteriol.* **136**:693–699.
52. Nesvizhskii, A. I., A. Keller, E. Kolker, and R. Aebersold. 2003. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**:4646–4658.
53. Niu, Y. D., et al. 2009. Host range and lytic capability of four bacteriophages against bovine and clinical human isolates of Shiga toxin-producing *Escherichia coli* O157:H7. *J. Appl. Microbiol.* **107**:646–656.
54. Niu, Y. D., et al. 2008. Comparison of fecal versus rectoanal mucosal swab sampling for detecting *Escherichia coli* O157:H7 in experimentally inoculated cattle used in assessing bacteriophage as a mitigation strategy. *J. Food Prot.* **71**:691–698.
55. Nolan, J. M., V. Petrov, C. Bertrand, H. M. Krisch, and J. D. Karam. 2006. Genetic diversity among five T4-like bacteriophages. *Virol. J.* **3**:30.
56. O’Flynn, G., R. P. Ross, G. F. Fitzgerald, and A. Coffey. 2004. Evaluation of a cocktail of three bacteriophages for biocontrol of *Escherichia coli* O157:H7. *Appl. Environ. Microbiol.* **70**:3417–3424.
57. Oliver, D. B., and R. A. Crowther. 1981. DNA sequence of the tail fibre genes 36 and 37 of bacteriophage T4. *J. Mol. Biol.* **153**:545–568.
58. Panos, G. Z., G. I. Betsi, and M. E. Falagas. 2006. Systematic review: are antibiotics detrimental or beneficial for the treatment of patients with *Escherichia coli* O157:H7 infection? *Aliment. Pharmacol. Ther.* **24**:731–742.
59. Pedrioli, P. G., et al. 2004. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **22**:1459–1466.
60. Pedulla, M. L., et al. 2003. Origins of highly mosaic mycobacteriophage genomes. *Cell* **113**:171–182.
61. Perry, L. L., et al. 2009. Sequence analysis of *Escherichia coli* O157:H7 bacteriophage PhiV10 and identification of a phage-encoded immunity protein that modifies the O157 antigen. *FEMS Microbiol. Lett.* **292**:182–186.
62. Petrov, V. M., et al. 2006. Plasticity of the gene functions for DNA replication in the T4-like phages. *J. Mol. Biol.* **361**:46–68.
63. Petrov, V. M., S. Ratnayaka, J. M. Nolan, E. S. Miller, and J. D. Karam.

2010. Genomes of the T4-related bacteriophages as windows on microbial genome evolution. *Virol. J.* **7**:292.
64. Plunkett, G., III, D. J. Rose, T. J. Durfee, and F. R. Blattner. 1999. Sequence of Shiga toxin 2 phage 933W from *Escherichia coli* O157:H7: Shiga toxin as a phage late-gene product. *J. Bacteriol.* **181**:1767–1778.
65. Rangel, J. M., P. H. Sparling, C. Crowe, P. M. Griffin, and D. L. Swerdlow. 2005. Epidemiology of *Escherichia coli* O157:H7 outbreaks, United States, 1982–2002. *Emerg. Infect. Dis.* **11**:603–609.
66. Raya, R. R., et al. 2006. Isolation and characterization of a new T-even bacteriophage, CEV1, and determination of its potential to reduce *Escherichia coli* O157:H7 levels in sheep. *Appl. Environ. Microbiol.* **72**:6405–6410.
67. Repoila, F., F. Tetart, J. Y. Bouet, and H. M. Krisch. 1994. Genomic polymorphism in the T-even bacteriophages. *EMBO J.* **13**:4181–4192.
68. Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**:276–277.
69. Rohrer, H., W. Zillig, and R. Mailhammer. 1975. ADP-ribosylation of DNA-dependent RNA polymerase of *Escherichia coli* by an NAD⁺:protein ADP-ribosyltransferase from bacteriophage T4. *Eur. J. Biochem.* **60**:227–238.
70. Ronner, A. B., and D. O. Cliver. 1990. Isolation and characterization of a coliphage specific for *Escherichia coli* O157:H7. *J. Food Prot.* **53**:944–947.
71. Sato, T., et al. 2003. Distinctiveness of the genomic sequence of Shiga toxin 2-converting phage isolated from *Escherichia coli* O157:H7 Okayama strain as compared to other Shiga toxin 2-converting phages. *Gene* **309**:35–48.
72. Sato, T., et al. 2003. Genome analysis of a novel Shiga toxin 1 (Stx1)-converting phage which is closely related to Stx2-converting phages but not to other Stx1-converting phages. *J. Bacteriol.* **185**:3966–3971.
73. Scholl, D., and C. Merrill. 2005. The genome of bacteriophage K1F, a T7-like phage that has acquired the ability to replicate on K1 strains of *Escherichia coli*. *J. Bacteriol.* **187**:8499–8503.
74. Shaikh, N., and P. I. Tarr. 2003. *Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications. *J. Bacteriol.* **185**:3596–3605.
75. Shub, D. A., H. Goodrich-Blair, and S. R. Eddy. 1994. Amino acid sequence motif of group I intron endonucleases is conserved in open reading frames of group II introns. *Trends Biochem. Sci.* **19**:402–404.
76. Sonnhammer, E. L., and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**:GC1–GC10.
77. Su, L. K., et al. 2010. Lysogenic infection of a Shiga toxin 2-converting bacteriophage changes host gene expression, enhances host acid resistance and motility. *Mol. Biol. (Moscow)* **44**:60–73.
78. Sullivan, M. B., M. L. Coleman, P. Weigle, F. Rohwer, and S. W. Chisholm. 2005. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* **3**:e144.
79. Sullivan, M. B., et al. 2010. Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. *Environ. Microbiol.* **12**:3035–3056.
80. Tetart, F., C. Desplats, and H. M. Krisch. 1998. Genome plasticity in the distal tail fiber locus of the T-even bacteriophage: recombination between conserved motifs swaps adhesin specificity. *J. Mol. Biol.* **282**:543–556.
81. Tetart, F., et al. 2001. Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J. Bacteriol.* **183**:358–366.
82. Tock, M. R., and D. T. Dryden. 2005. The biology of restriction and anti-restriction. *Curr. Opin. Microbiol.* **8**:466–472.
83. Vakharia, H., and R. Misra. 1996. A genetic approach for analysing surface-exposed regions of the OmpC protein of *Escherichia coli* K-12. *Mol. Microbiol.* **19**:881–889.
84. Van Sinderen, D., et al. 1996. Sequence analysis and molecular characterization of the temperate lactococcal bacteriophage rlt. *Mol. Microbiol.* **19**:1343–1355.
85. Villegas, A., et al. 2009. The genome and proteome of a virulent *Escherichia coli* O157:H7 bacteriophage closely resembling *Salmonella* phage Felix O1. *Virol. J.* **6**:41.
86. Wilson, J. H. 1973. Function of the bacteriophage T4 transfer RNA's. *J. Mol. Biol.* **74**:753–757.
87. Yates, J. R., III, J. K. Eng, A. L. McCormack, and D. Schieltz. 1995. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67**:1426–1436.
88. Young, I., I. Wang, and W. D. Roof. 2000. Phages will out: strategies of host cell lysis. *Trends Microbiol.* **8**:120–128.
89. Yu, S. L., et al. 1998. Characterization of a phage specific to hemorrhagic *Escherichia coli* O157:H7 and disclosure of variations in host outer membrane protein OmpC. *J. Biomed. Sci.* **5**:370–382.
90. Yu, S. L., K. L. Ko, C. S. Chen, Y. C. Chang, and W. J. Syu. 2000. Characterization of the distal tail fiber locus and determination of the receptor for phage AR1, which specifically infects *Escherichia coli* O157:H7. *J. Bacteriol.* **182**:5962–5968.
91. Zdobnov, E. M., and R. Apweiler. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**:847–848.
92. Zuber, S., et al. 2007. Genome analysis of phage JS98 defines a fourth major subgroup of T4-like phages in *Escherichia coli*. *J. Bacteriol.* **189**:8206–8214.