

## Genome Analysis of Phage JS98 Defines a Fourth Major Subgroup of T4-Like Phages in *Escherichia coli*<sup>∇†</sup>

Sophie Zuber, Catherine Ngom-Bru, Caroline Barretto, Anne Bruttin,  
Harald Brüssow,\* and Emmanuel Denou

Nestlé Research Center, Nestec Ltd., P.O. Box 44, CH-1000 Lausanne 26, Switzerland

Received 30 May 2007/Accepted 4 August 2007

Numerous T4-like *Escherichia coli* phages were isolated from human stool and environmental wastewater samples in Bangladesh and Switzerland. The sequences of the major head gene (*g23*) revealed that these coliphages could be placed into four subgroups, represented by the phages T4, RB69, RB49, and JS98. Thus, JS98 defines a new major subgroup of *E. coli* T4-like phages. We conducted an analysis of the 169-kb JS98 genome sequence. Overall, 198 of the 266 JS98 open reading frames (ORFs) shared amino acid sequence identity with the reference T4 phage, 41 shared identity with other T4-like phages, and 27 ORFs lacked any database matches. Genes on the plus strand encoded virion proteins, which showed moderate to high sequence identity with T4 proteins. The right genome half of JS98 showed a higher degree of sequence conservation with T4 and RB69, even for the nonstructural genes, than did the left genome half, containing exclusively non-structural genes. Most of the JS98-specific genes were found in the left genome half. Two came as a hyper-variability cluster, but most represented isolated genes, suggesting that they were acquired separately in multiple acquisition events. No evidence for DNA exchange between JS98 phage and the *E. coli* host genome or coliphages other than T4 was observed. No undesired genes which could compromise its medical use were detected in the JS98 genome sequence.

Research on the *Escherichia coli* bacteriophage T4 started in the 1940s and became a cornerstone of molecular biology. Decades ago, phage T4 combined just the right blend of genetic complexity and experimental tractability to make it a major model system in biology (18). Technological progress now allows molecular biologists to work with complex eukaryotic model systems, and when genomics became popular, it started with bacteria, not phages. Interest in comparative T4 genomics is relatively recent and led to the compilation of a number of T4-like phage genomes in the Tulane T4 phage database (<http://phage.bioc.tulane.edu/>). Fascinating results were achieved by sequence analysis of distant relatives of T4 infecting cyanobacteria (15, 21, 33). Based on sequence analysis of the major head gene, T4 phages were classified as T-even and pseudo-, schizo-, and exo-T-even phages (36). Surprisingly, the genomic similarity and diversity within the T-even group are less well documented (27, 38), although important insights into the genome evolution of T4 phages can be expected from such comparisons (11). We therefore decided to conduct comparative genomic analyses within that group.

There is also a medical interest in gaining a better knowledge of *E. coli* phages closely related to T4. *E. coli* is a versatile pathogen causing urinary and gastrointestinal infections. The diarrhea burden is especially large for children from develop-

ing countries (2, 4). Diarrhea represents the second most frequent cause of morbidity and mortality (32), and *E. coli* is responsible for one-third of cases (2). The use of oral rehydration solution has substantially reduced mortality (3), but its application does not treat or prevent infections. Vaccines against *E. coli* diarrhea are not yet available (29, 31), and antibiotics are of limited use (14, 23). Considering these issues, it is not surprising that the old idea of phage therapy was taken up against *E. coli* diarrhea (5). However, the reference T4 phage has only a narrow host range on pathogenic *E. coli* strains. Therefore, we had to isolate phages with a broader host range from sewage and stool samples of children hospitalized with diarrhea (10). Field studies in Bangladesh identified JS98-like phages as frequent isolates. Partial sequencing of its genome revealed JS98 to be a distant relative of T4, suggesting a hitherto uncharacterized new branch of T-even phages (9). Closely related phages have also been isolated from other geographical areas (16a). To achieve optimal coverage of pathogenic *E. coli* strains, T4-, RB69-, JS98-, and RB49-like phages are part of our phage cocktail. Since all of these phage groups except for JS98 are represented by at least one complete genome sequence, we targeted JS98 for sequencing. Due to the presence of many host-lethal genes, we could not obtain the complete phage genome sequence of JS98 with techniques based on phage DNA cloning (9). Here we report that the newer pyrosequencing technology yielded the complete phage JS98 sequence at a fraction of the time and cost of previously used sequencing approaches. In analyzing this sequence, we asked the following two questions. Does phage JS98 contain genes that represent a potential safety concern for oral application in humans, and what does comparative genomics of JS98 tell us about genetic diversification and evolution of T-even phages?

\* Corresponding author. Mailing address: Nestlé Research Centre, Nutrition and Health Department/Food and Health Microbiology, Vers-chez-les-Blanc, CH-1000 Lausanne 26, Switzerland. Phone: 41 21 785 8676. Fax: 41 21 785 8544. E-mail: harald.brussow@rdls.nestle.com.

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

∇ Published ahead of print on 10 August 2007.

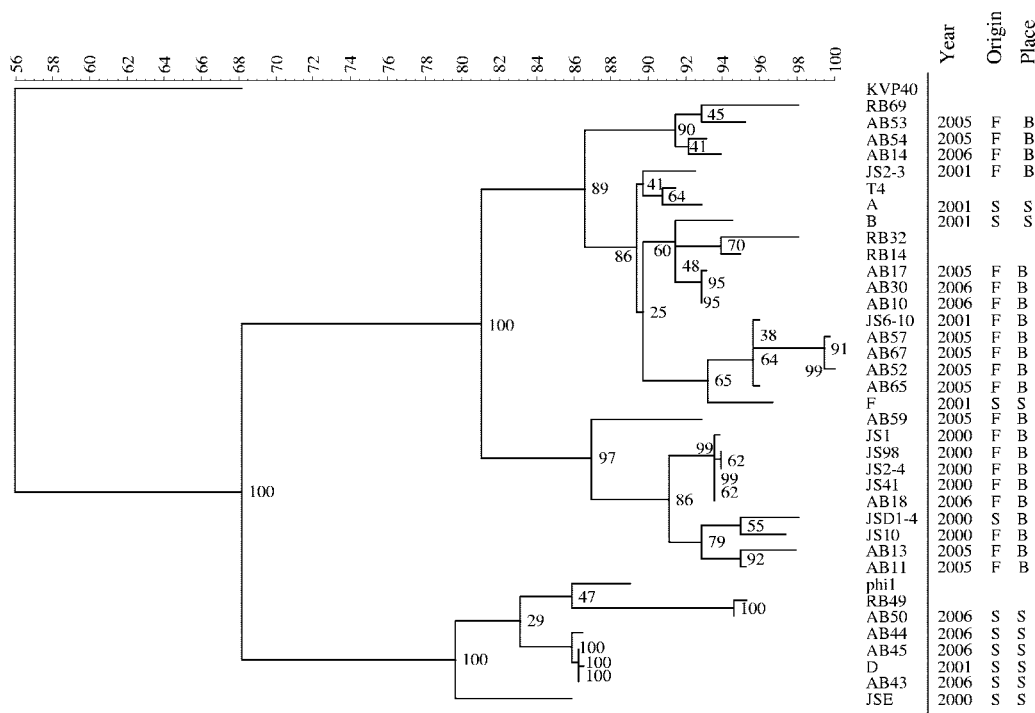


FIG. 1. *g23* tree analysis. Major head gene *g23* tree analysis was performed with the sequenced PCR products from our field isolates of T4-like phages. The tree is based on nucleotide sequence alignments (corresponding to codons 346 to 1118 in T4). Origins (S, sewage water; and F, fecal content), places (B, Bangladesh; and S, Switzerland), and years of isolation are indicated for the phages from our survey. The numbers at the nodes give the bootstrap probabilities, and the scale above gives percentages of base pair sequence identity. The tree was rooted with the *Vibrio* T4 phage KVP40 genome sequence. Their codes are indicated at the level of the twigs of the tree.

**MATERIALS AND METHODS**

**Phylogenetic tree.** The different *g23* sequences were amplified by PCR, using the primers Mzia1 and CAP.8 (36), and the PCR products were sequenced (Fasteris SA, Geneva, Switzerland). Alternatively, the sequences were retrieved from the published genomes. Raw sequence data were transferred into BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium), where a consensus sequence was determined. A similarity matrix and phylogenetic trees were created based on the maximum parsimony and neighbor-joining methods. The reliability of the groups was evaluated by bootstrap analysis with 500 resamplings.

**Sequencing.** T4-like phage strain JS98 was isolated from the stool of a pediatric diarrhea patient in Bangladesh (10) and grown on an *E. coli* K-12 strain. Phage DNA was purified and sent to 454 Life Sciences (Branford, CT) for commercial sequencing. There the phage DNA was amplified by an emulsion-based method, sequenced by synthesis using a pyrosequencing protocol (22), and assembled de novo into a single contig.

**Bioinformatics analysis and annotation of genome sequence data.** Genome sequence comparisons were carried out using the MUMmer package (version 3.18) (19; <http://www.tigr.org/software/mummer>) and the EMBOSS (The European Molecular Biology Open Software Suite) software STRETCHER, with the parameters set at default values (26). We used Artemis software as a sequence viewer and annotation tool (30). Open reading frames (ORFs) were predicted using Glimmer software (version 3.02) (12) and based on nucleotide and amino acid sequence alignment searches (BlastN, TblastN, BlastX, and BlastP), using the T4-like genome database available through the Tulane website (<http://phage.bioc.tulane.edu>) and the nonredundant database from NCBI. The basic prerequisites for an ORF were the presence of one of the three potential start codons, i.e., ATG, TTG, or GTG, and a length of at least 25 encoded amino acids. A search for tRNA genes was done with the tRNAscan-SE program (version 1.23) (20). Homology assignments between genes from other T4-like phages and predicted ORFs of phage JS98 were based on amino acid sequence alignment searches (BlastP) and were accepted only if the statistical significance of the sequence similarities (E value) was  $\leq 0.001$ , the bit score was  $\geq 50$ , and the percent identity between the aligned sequences was  $\geq 30\%$ . Functional annotations were based on the homology assignments with the T4 genes, and functional

classifications were performed using the COG (clusters of orthologous groups of proteins) database (34).

**Safety evaluation.** We constructed a database of undesirable genes (DUG) by searching public databases available at NCBI for a list of antibiotics and virulence terms. BlastP and BlastN searches were performed against the DUG and against the 15 *E. coli* genomic sequences currently available at NCBI. Only hits with E values of  $\leq 0.01$  (BlastP and BlastN) and bit scores of  $\geq 50$  (BlastP) were considered to constitute significant matches. Searches for specific protein domains and conserved motifs with known function were performed using the Conserved Domain Architecture Retrieval Tool (CDART) available at NCBI and the InterPro (<http://www.ebi.ac.uk/interpro/>) databases. The 266 predicted protein sequences of JS98 were screened for similarities to currently known protein food allergens by comparing them to the amino acid sequences present in the Food Allergy Research and Resource Program allergen database available at <http://www.allergenonline.com>. Only hits with E values of  $\leq 0.01$  and bit scores of  $\geq 50$  were considered to constitute significant matches.

**Nucleotide sequence accession number.** The sequence data were deposited at GenBank under accession number EF469154, which replaces accession numbers AY746495, AY746496, and AY746497.

**RESULTS**

**Phage subgroups.** To obtain an overview about the genetic diversity of the T4-like phages isolated in our ecological survey (10), we sequenced the amplification products from a diagnostic PCR (36). The test was based on the *g23* sequence encoding the major head protein. We did this analysis on a subset of 31 of the 60 isolated phages. A tree analysis of these *g23* sequences revealed four subgroups of the T4-like phages in our collection (Fig. 1). Phages related to RB49, a representative pseudo-T-even phage (13), were the dominant isolates from

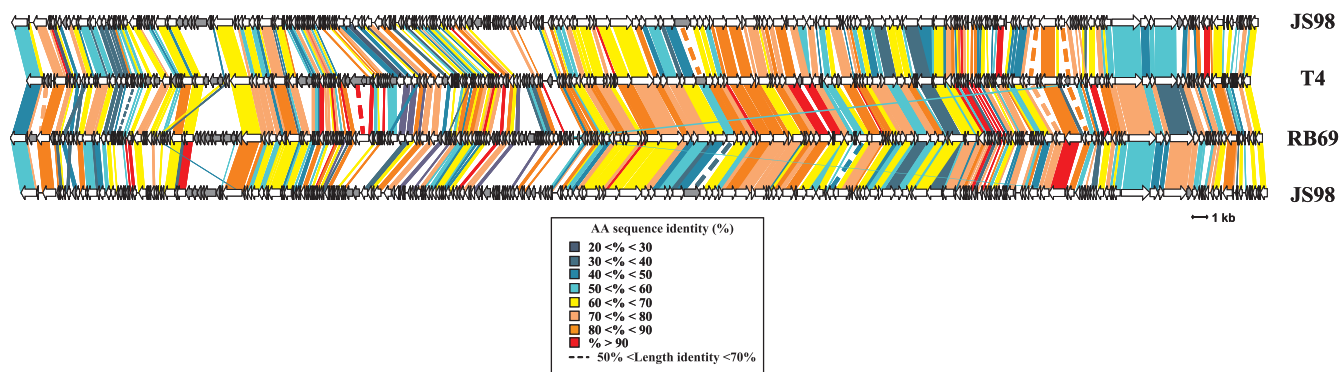


FIG. 2. Alignment of genome maps of phages JS98, T4, and RB69. The ORFs are indicated by arrows as follows: white arrows indicate ORFs which share amino acid sequence identity over  $>70\%$  of the sequence length, and gray arrows indicate ORFs that are specific to the indicated phage genome in this three-phage comparison. ORFs sharing amino acid sequence identity are linked by color shading; the color scale at the bottom of the figure indicates the percentage of amino acid sequence identity between the compared predicted proteins. The JS98 map is shown at the top and bottom to allow comparisons with both T4 and RB69.

Swiss environmental water samples (six of nine isolates). In our survey, no RB49-like phages were isolated from stool and water samples from Dhaka, Bangladesh. All Bangladeshi stool isolates belonged to the T-even phages. Tree analysis suggested that this group can be separated into branches represented by RB69, T4, and JS98. Nine, nine, and three Bangladeshi stool isolates clustered with the JS98, T4, and RB69 branches, respectively.

**Sequencing.** Sequencing of the JS98 genome was initially conducted by shearing phage JS98 DNA into 1-kb fragments and cloning them into a plasmid vector. The inserts were sequenced to eightfold coverage. We did not obtain a complete genome sequence but obtained 21 contigs. By projecting them on a T4 genome map, we were able to close a few gaps by sequencing PCR products linking adjacent contigs. By this approach, we obtained 55% of the genome connected in three contigs (see Fig. S1 in the supplemental material). However, many regions were not represented in the assembly, probably because they contained host-lethal genes (9). To circumvent this problem, we sent the phage DNA to 454 Life Sciences, which employed the newer pyrosequencing technique (22). Overall, 61,936 reads were assembled into a single contig of 170,523 bp for JS98, which is slightly larger than the genome of T4 (168,903 bp). Since T4 phage genomes do not contain repeat regions, the assembly did not present a problem. The correctness of the assembly was verified by comparing an alignment of the JS98 genome with the T4 genome at the DNA (see Fig. S2 in the supplemental material) and protein (see Fig. S3 in the supplemental material) sequence levels. Most reads showed lengths of between 100 and 140 bp, yielding a sharp peak around 110 bp (see Fig. S4 in the supplemental material). The coverage of the sequencing varied from 10- to 70-fold; most regions showed coverage of at least 30-fold (see Fig. S5 in the supplemental material). In regions with  $>30$ -fold coverage, a 30-bp overlap of at least 10 readings was achieved (see Fig. S6 in the supplemental material). Even the poorest cases of coverage allowed a definitive assembly (see Fig. S7 in the supplemental material). The two JS98 sequences could be compared over 94,462 bp and revealed 99.97% sequence identity. Sequence discrepancies were resolved mostly in favor of the newer sequencing method (data not shown).

**JS98 comparison with RB69 phage.** Alignment of the JS98 genome with T4-like phages in the Tulane database showed that JS98 was most closely related to phage RB69. In both DNA sequence and protein sequence dot plots, we observed a frequently interrupted but straight diagonal line between both phages (see Fig. S2 and S3 in the supplemental material). Overall, the two genomes are colinear but are frequently interrupted by replacements with unrelated genome segments of comparable lengths.

Figure 2 shows an alignment of the JS98 and RB69 genome maps. The protein sequence relatedness between them is color-coded. The right genome halves are closely related. They cover two rightward-oriented structural gene clusters and, between them, a cluster of leftward-oriented nonstructural genes, mainly encoding proteins involved in nucleotide metabolism. The degree of sequence identity varied substantially: it was highest for the head and nucleotide metabolism genes and lowest for three groups of structural genes. The degree of sequence conservation thus did not follow the structural/nonstructural gene division.

JS98 and RB69 share highly related head and tail genes but use distinct base plate and tail fiber genes. Notably, the differences are located over those base plate genes that encode the tail fiber socket and the interacting proximal tail fiber genes, suggesting variation for interacting but chromosomally separated genes. The distal tail fiber genes were again closely related between JS98 and RB69.

The left genome halves cover exclusively leftward-oriented nonstructural genes, including a large DNA replication module. They were substantially less well aligned than the right genome halves. Large regions of nonaligned genome segments showed mainly gene replacements. Only a few gene inversions affecting a single gene were observed.

**JS98 comparison with T4.** After RB69, phage T4 showed the next best dot plot alignment with JS98 (see Fig. S2 and S3 in the supplemental material) from the entries in the NCBI database. Figure 2 shows a comparison of the two genome maps. Overall, rather similar observations were made to those described above in the JS98-RB69 comparison. The major new observation was the lack of high sequence identity between the distal tail fiber genes of T4 and JS98. Furthermore, regions of

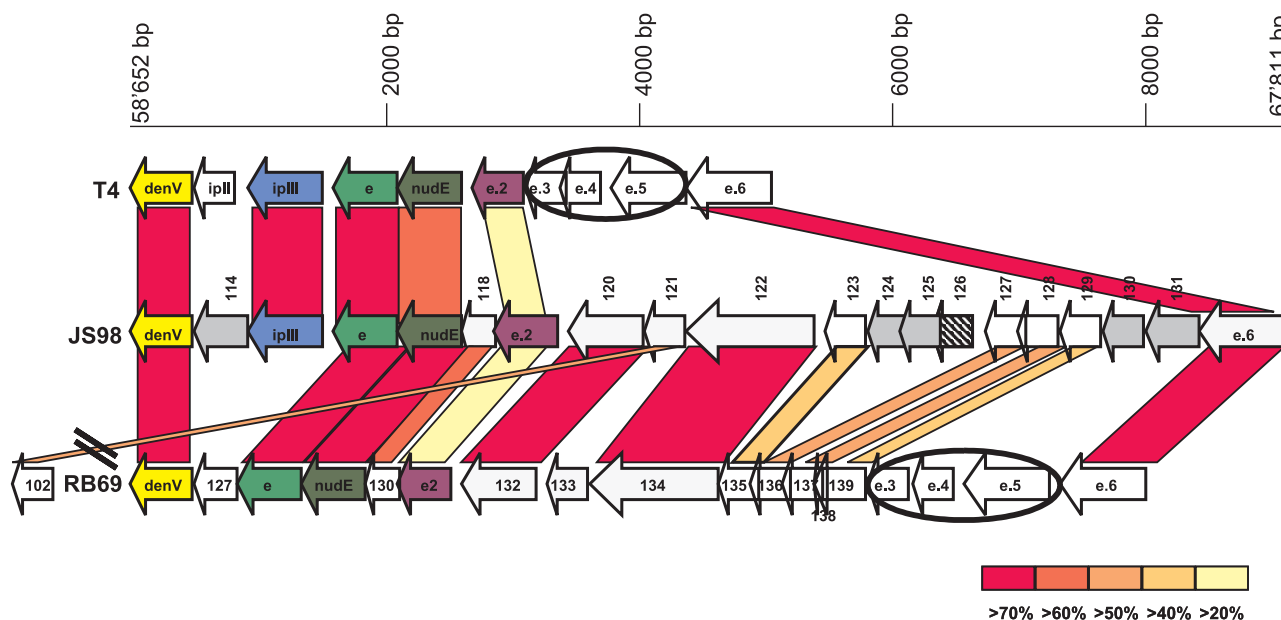


FIG. 3. Alignment and comparison of a 10-kb region showing a high degree of variability between the phages T4 (top), JS98 (middle), and RB69 (bottom). Genes are colored according to their T4 functional assignments (25). Gray indicates genes unique to JS98. Gene 126 is shown hatched, as it shows 40.5% homology to ORF 063 of phage 44RR. The T4 ORFs are annotated with their conventional gene names, JS98 ORFs are numbered or named after the corresponding T4 homologues, and RB69 genes are quoted with the annotations given to them in the GenBank entry (accession number NC\_004928). Amino acid sequence identities between genes were determined using STRETCHER and are indicated by connections of red to yellow shading, according to the color key provided at the bottom right. The black ovals indicate homology between T4 and RB69. The top line provides a base pair scale and the positions of the first and last depicted JS98 genes.

gene replacement detected in the left genome halves were no longer of comparable lengths; in two cases, genes on the JS98 map lacked a complement in T4. One case of such a complex gene replacement is illustrated in Fig. 3. Over this region, JS98 shared greater similarity with RB69 than with T4. T4 and RB69 shared genes lacking in JS98, and JS98 contained genes not found in either RB69 or T4. One could explain the observed gene constellation by a combination of insertion events (e.g., ORFs 124 to 126 in JS98), gene replacements (e.g., JS98 ORFs 130 to 131 versus RB69 *e.3* to *e.5*), and DNA rearrangements (JS98 ORF 121 versus RB69 ORF 102).

Phages T4 and JS98 showed average GC contents of 35.3 and 39.5%, respectively, which are much lower than that of their *E. coli* host, with a 50.8% GC content. Only two phage genome regions demonstrated a significant deviation from the average GC content (see Fig. S8 in the supplemental material). One is centered at the major head gene *g23*, where JS98 and T4 showed local GC contents of 47.6% and 45%, respectively. The other is located over the tail fiber cluster, a region where lateral gene transfer in this phage group is known to be frequent (35).

**RB69 comparison with T4.** According to the *g23* tree analysis, RB69 and T4 belong to separate branches of T4 coliphages, but these branches are more closely related to each other than any of them is to JS98 (Fig. 1). Comparative genomics confirmed this relationship: phages T4 and RB69 showed a substantially larger number of genes sharing >80% sequence identity than did the JS98-T4 or JS98-RB69 alignment (Fig. 2). The sequence identity between T4 and RB69 was especially marked over the right genome halves. The left genome halves

varied more substantially, as gene replacements and insertions/deletions were also frequently observed between RB69 and T4.

**JS98 genome map. (i) Plus strand.** We next investigated the JS98 genome in greater detail. Annotation of the JS98 genome revealed 266 ORFs. Notably, 198 of the 266 JS98 ORFs (74% of the total) shared significant amino acid sequence identity with T4 proteins. Based on their similarity with biologically defined T4 proteins, 114 ORFs of JS98 (43% of the total) could be given a functional annotation (Fig. 4).

Following the T4 convention, the gene *rIIA* was positioned at the start of the map. The plus strand exclusively encodes the putative virion structural proteins sharing substantial sequence identity with T4 virion proteins (e.g., 68.8% amino acid sequence identity between T4 Gp6 and its homologue in JS98). Structural genes defined three separate clusters of rightward-transcribed genes, including an about 30-kb cluster of base plate wedge/head/tail genes (genes 53 to 24), a smaller base plate hub gene cluster in its vicinity (genes 51 to 54), and further away, a tail fiber gene cluster consisting of a few large genes (genes 34 to *t*) (9) (Fig. 4).

**(ii) Minus strand.** The longest cluster of leftward-transcribed genes (i.e., genes carried on the minus strand) extends from a T4 *g4* homologue to *asiA* homologues covering the first half of the genome (and a smaller part of the opposite end, a consequence of the artificial cutting of the genome between *rIIA* and *-B*). Database homologies in this region defined a large cluster of DNA replication genes. In addition, several nucleotide metabolism, translation, and transcription regulation genes were identified (see the color code for the ORFs in Fig. 4). A second large cluster of leftward-transcribed genes

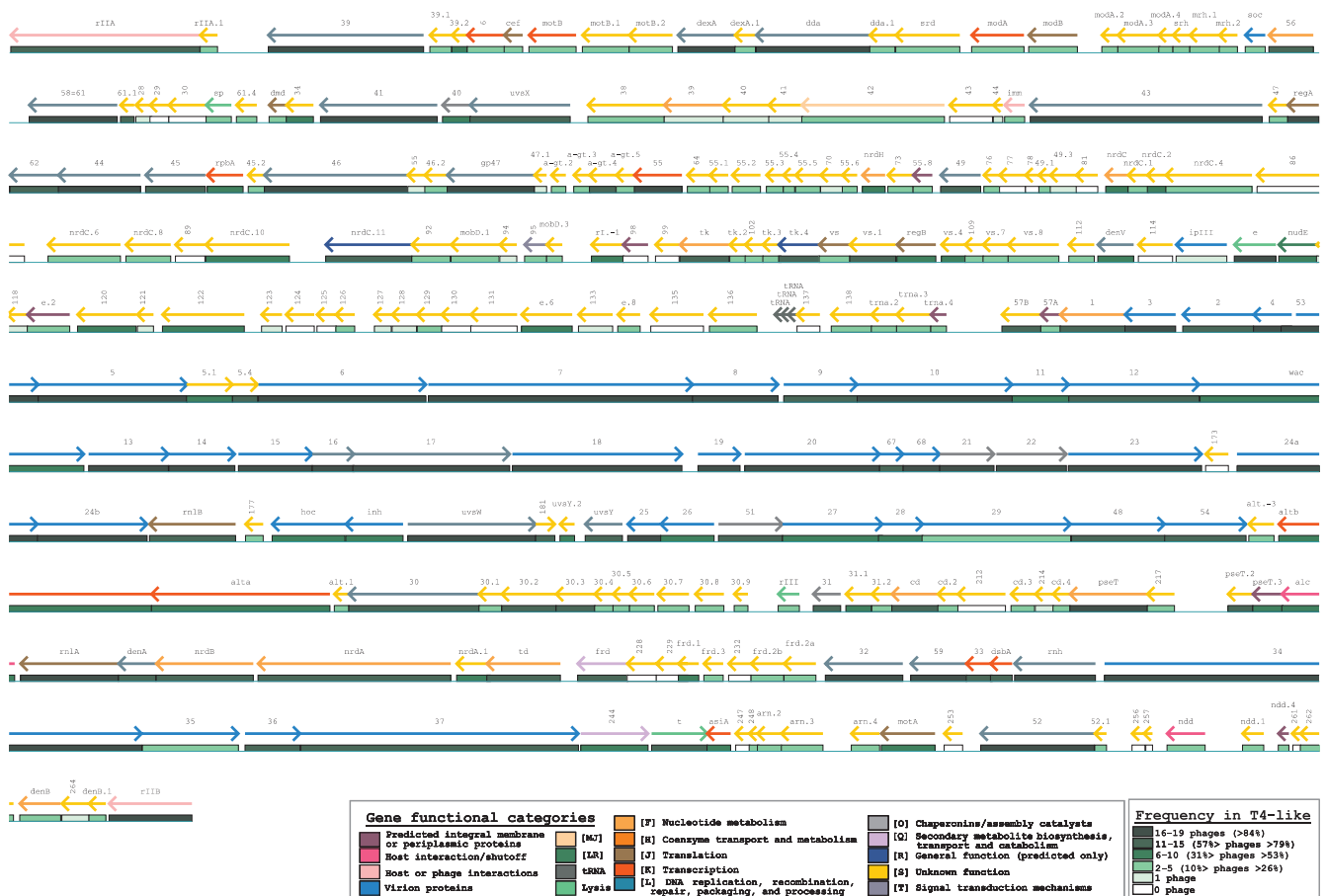


FIG. 4. Annotated genome map of bacteriophage JS98. Following convention, the map starts at the top left with the *rIIA* gene and ends at the bottom left with the *rIIB* gene. The genome was divided into 15-kb segments that are to be read from left to right and from top to bottom. The individual ORFs are depicted as arrows, with the orientation of the arrows indicating whether the genes are carried on the Watson or Crick strand. The color of the arrow identifies the functional category into which the homologous T4 gene was classified (25). The color code for gene function and the COG letter code are provided in the bottom center frame of the figure. The 266 predicted JS98 ORFs plus three tRNA genes are annotated above the corresponding arrows with the names of the homologous T4 genes (letter code or horizontal number code). If the JS98 ORF lacks a T4 gene complement, we attributed a number to the gene, starting from *rIIA*. To distinguish these JS98 ORFs from T4 genes which also carry numbers as gene identifiers (horizontal numbers), we annotated the JS98 genes lacking a T4 homologue with vertical numbers. The first such JS98 ORF without a T4 complement is found in the second line, following the T4 homologue *uvrX*, and is annotated as JS98 ORF 38. Three tRNA genes are indicated with dark olive arrows, located between JS98 ORFs 136 and 137. Below the arrows are boxes whose colors indicate how many phages from the Tulane database contain a protein which shares protein sequence identity with the JS98 gene. The key to the color code for the prevalence of the JS98 genes is provided in the bottom right frame.

extends over nearly 30 kb, from a T4 homologue for the RNase H gene (*rnh*) to the T4 adenosyl-ribosyltransferase homologue *alt* gene, located between the base plate hub and the tail fiber gene clusters. This region contains several nucleotide metabolism, DNA replication, and transcription genes.

**Comparison with proteins in the Tulane T4 phage database.** We searched the Tulane protein database of T4-like phage genomes with the JS98 sequence, using TblastN searches, and determined the number of T4-like phages with a homologous ORF for each JS98 ORF (Fig. 4). Significantly, 45 of the 68 ORFs from JS98 that lacked a T4 match shared protein sequence identity with another T4-like phage (see Table S1 in the supplemental material). Figure 4 depicts the degree of conservation for each JS98 ORF, expressed in a color code. Dark green indicates highly conserved genes with matches in more than 10 other T4-like phages. With few exceptions (polynucleotide kinase *pseT*, RNA ligase *rmlA*, and *rIIB* genes),

the highly shared core genes come from the DNA replication and virion structural modules. The major structural genes are located in a tight cluster of highly conserved genes. In contrast, the DNA replication genes are less well conserved and are often separated by nonconserved phage genes. Notably, all JS98 genes located on the plus strand shared sequence identity with the predicted proteins from many T4-like phages.

The genes on the minus strand showed, on average, much less representation in the Tulane T4 phage database (<http://phage.bioc.tulane.edu/>) than the genes on the plus strand. For example, all genes lacking either NCBI or Tulane database matches (26 ORFs) are located on the minus strand (depicted with white boxes under white arrows in Fig. 4). These JS98-specific genes are not clustered on the genome: they occur either as single genes or as two adjacent genes (Fig. 2).

**Hypothetical and undesired proteins.** Table S1 in the supplemental material provides a list of the 68 ORFs from JS98

TABLE 1. Phage JS98 ORFs showing homology to *E. coli* ORFs<sup>a</sup>

JS98 ORF	T4 homologue	Hit against NCBI database of <i>E. coli</i> genomes			% ID	Score	Accession no.
		<i>E. coli</i> strain	Gene or locus tag	Protein			
003*	39	UTI89	<i>gyrB</i>	DNA gyrase subunit B	28	164	YP_543207
011	<i>dexA</i>	CFT073	<i>recE</i>	Exodeoxyribonuclease VIII	25	55	NP_753311
013*	<i>dda</i>	K12	<i>recD</i>	Exonuclease V (RecBCD complex), alpha chain	31	54	NP_417296
037*	<i>uvsX</i>	K12	<i>recA</i>	Recombinase A	24	62	NP_417179
046*	43	E22	EcolE2_01003982	COG0417; DNA polymerase elongation subunit	26	64	ZP_00727779
100*	<i>tk</i>	K12	<i>tdk</i>	COG1435; thymidine kinase	51	192	NP_415754
159*	<i>wac</i>	E22	EcolE2_01000575	COG1061; DNA or RNA helicase of superfamily II	21	53	ZP_00730960
180*	<i>uvsW</i>	E22	EcolE2_01000575	COG1061; DNA or RNA helicase of superfamily II	21	59	ZP_00730960
223*	<i>nrdB</i>	O157:H7 EDL933	<i>nrdB</i>	Ribonucleotide-diphosphate reductase beta subunit	54	396	NP_288809
224*	<i>nrdA</i>	K-12	<i>nrdA</i>	Ribonucleotide-diphosphate reductase alpha subunit	54	763	NP_416737
226*	<i>td</i>	HS	EcolH_01002607	COG0207; thymidylate synthase	47	256	ZP_00706865
254*	52	B7A	EcolB7_01002444	COG0188; type 2A topoisomerase (DNA gyrase/topoisomerase II, topoisomerase IV), A subunit	23	109	ZP_00715856
095*		K-12	<i>ybil</i>	COG1734; DnaK suppressor protein	41	52	NP_415324
158*	12	101-1	Ecol1_01003560	COG5301; phage-related tail fiber protein	27	51	ZP_00923911
240*	34	O157:H7 strain Sakai	Ecs0542	Hypothetical protein	20	56	NP_308569
243*	37	E22	EcolE2_01000786	COG5301; phage-related tail fiber protein	29	82	ZP_00730675
244	38	HS	EcolH_01000816	COG5301; phage-related tail fiber protein	33	53	ZP_00705098
151*	5.4	UTI89	UTI89_C1675	COG4104; uncharacterized conserved protein (PAAR motif)	36	51	YP_540684

<sup>a</sup> BlastP results (E values of <0.01 and scores of >50) for JS98 proteins listed as queries against the 15 *E. coli* genomes present in the NCBI database (complete genomic sequences and whole-genome shotgun sequences) are listed. The T4 homologues of the JS98 ORFs are indicated in the second column. Proteins sharing a low degree of homology are shown in parentheses. \*, hit against the DUG.

that lacked any sequence homology with T4. Some clustering of these non-T4-related genes was observed on the JS98 genome. The hypothetical and no-hit proteins from JS98, which lacked matches with the Tulane database, were screened using InterProScan. A PROSITE motif (TonB-dependent receptor protein signature 1) was found for ORF253, and a PFAM domain (anticodon nuclease activator protein) was found for ORF257. Nine of the hypothetical proteins contained a signal peptide motif, and among these, six also showed one or two transmembrane domains. This small number of additional links demonstrates the seclusion of T4 phages from the entries in the database. We also screened all JS98 genes against our DUG and obtained no matches. Likewise, matches with protein food allergens in the FARRP Food Allergen Database (<http://www.allergenonline.com>) were not found.

**Links to *E. coli* genes.** Next, we screened for possible horizontal gene transfer by comparing the JS98 genome to all available *E. coli* genome sequences. Eighteen predicted JS98 proteins shared sequence identity with *E. coli* proteins (Table 1). The first 12 proteins on this list belong to the category of DNA replication and DNA transaction genes. Notable are the NrdA and NrdB proteins, which are the  $\alpha$  and  $\beta$  subunits of ribonucleotide reductase, and the thymidylate synthetase Td, which shared 54 and 47% amino acid identity, respectively, with their *E. coli* homologues. In phage T4, these adjacent genes are interrupted or flanked by mobile DNA elements (introns and intron-homing endonucleases) (17). An amino acid identity of 51% was also found for the cellular and viral thymidine kinases. Interestingly, in T4 the *tk* gene is also fol-

lowed by mobile DNA elements. However, no DNA sequence identity was detected between these viral and *E. coli* proteins, arguing against a recent horizontal gene transfer event as an explanation.

Table 1 also lists low-grade similarities with three *E. coli* proteins that most likely represent tail fiber genes from prophage remnants. They occur at a JS98 map position where T4 encodes tail fibers (*g12*, short tail fiber; *g37*, large distal tail fiber; and *g38*, tail fiber assembly catalyst). All of these genes also have many matches to the T4 phage family (Fig. 4).

## DISCUSSION

What can be learned from the JS98 sequence with respect to the phage JS98 safety profile and its use in human volunteers? The phage JS98 genome analysis did not reveal harmful genes encoding virulence factors, genes encoding proteins that alter the antigenicity of the bacterial host, or antibiotic resistance genes. The absence of such genes is not a trivial observation, since genes conferring virulence traits to pathogenic *E. coli* are frequent findings in the genomes of lambda-like *E. coli* phages (6, 8). The lack of pathogenicity genes in T4 phages was anticipated, as none have been reported during decades of intensive T4 phage research. Furthermore, human volunteers who received T4 phage orally in a safety trial showed no adverse effects (7). On overall balance, from the 266 JS98 ORFs, 198 had T4 homologues, a further 41 ORFs matched other T4-like phages, and only 27 ORFs lacked any database matches. A 10% rate of nonattributed genes is very low compared to those

of other phage genome analyses, where frequently the majority of the genes lacked database matches (24). The low percentage of nonattributed genes in phage JS98 likely reflects the intensive sequencing efforts within the T4 phage group (11, 28). The lack of recognizable protein motifs precludes any meaningful speculation about the function of the 10% of unknown genes in the JS98 genome. This situation is not specific to JS98, as T4 carries a similar percentage of nonattributed genes (25). If this number of phage strain-specific ORFs is typical, then the sum of variable T4-like genes will soon exceed the number of conserved core genes within the T4-like sequence space. Actually, there are relatively few T4 core functions, and despite decades of research, almost half of the T4 genes do not yet have an assigned function, with only 62 of the T4 genes being essential under standard laboratory conditions (25). While we cannot yet define the function of the strain-specific genes in the T4 group, these genes are unlikely candidates for host virulence genes. However, our ignorance regarding T4 is scientifically disturbing and probably reflects the fact that T4 has not yet been investigated in its natural environment, the gut.

A crucial aspect of any safety analysis of phage genomes concerns the degree of genetic exchange with their bacterial hosts. A total of 18 JS98 ORFs had sequence relatedness with *E. coli* proteins (Table 1). All showed, at the same time, sequence matches to T4-like phage proteins. The highest degree of sequence identity with *E. coli* proteins was 51 to 54% amino acid identity (*tk* and *nrdA/B*). In fact, orthologous genes for these functions are widely distributed, and a previous phylogenetic tree analysis suggested that, for example, the T4 thymidylate synthetase branched off before the split between the eukaryotic and bacterial orthologues (25). Obviously, such genes do not constitute evidence of horizontal gene transfer. The genetic isolation of JS98 from its *E. coli* host is further underlined by the drastically lower GC content of the phage DNA. Only two regions in the JS98 genome (and those of many other T4-like phages) showed a significantly higher GC content than the average. One region covers the major phage capsid gene and is therefore an unlikely candidate for lateral gene transfer from the bacterial host. Since Gp23 is one of the most abundantly expressed T4 proteins, the higher GC content may represent an adaptation to the codon usage of *E. coli* to optimize gene expression. The second locus is the distal part of the tail fiber gene cluster. This region is known to undergo gene shuffling for host range changes (35). As long as no known virulence genes are carried with the phage into a new species, this observation does not represent a safety concern, either.

Furthermore, we did not observe genes in JS98 whose best hits were with phages outside the T4 group. This observation suggests that gene exchange of T4 phages with other coliphages is rare. This lack of genetic exchange with both the bacterial host genome and other phage genomes may be explained partially by the rapid and complete degradation of the bacterial genome at an early infection stage (37). T4 phages may simply lack a utilizable source of foreign genes for gene transfers to occur to a reasonable extent.

JS98 and related phages were tested in mice, and no adverse events were observed (A. Bruttin et al., unpublished data). On the basis of the genome analysis and these animal safety tests,

JS98-like phages were then introduced into our phage cocktail for further safety evaluations.

What can be learned from the JS98 sequence analysis with respect to the evolution of the T4 genome? Based on the strikingly different GC content from that of its host, T4 has not evolved in *E. coli*. Recent structural analysis of the T4 capsid protein Gp24 (and, to a lesser extent, Gp23) revealed close conformational similarity with the major head protein from the lambdoid coliphage HK97 (16). This observation suggests a common ancestor for T4- and lambda-like tailed phages with respect to the assembly and structure of double-stranded DNA phage heads. However, the lack of any sequence similarity between these structurally related phage proteins suggests an ancestor deep in the evolutionary past, long before the evolution of *Enterobacteriaceae*. Some glimpses into the distant past of T4-like phages can be derived from comparisons of T4 with distantly related phages from cyanobacteria (15, 21, 33).

Information on genetic mechanisms, which are the motor for short-term T4 evolution, is better derived from comparisons of more closely related T4 phages. A detailed analysis of the pseudo-T-even coliphage RB49 with the reference T4 phage was published recently (11). The authors distinguished four segments of a conserved core genome, with two carrying virion genes and two carrying DNA replication genes. Replication and virion gene clusters were separated by hyperplastic regions containing mostly novel genes of unknown function and origin. Moving to more distant relatives of T4, such as the *Aeromonas* phage Aeh1, a similar pattern of conservation was observed. Differences were a lower degree of DNA sequence identity between Aeh1 and T4 than that between RB49 and T4, the splitting of the DNA replication genes over three gene clusters, and the larger sizes of the hyperplastic regions in Aeh1, which also contains a substantially larger genome (233 kb) than T4 (169 kb). The authors noted a gradient of decreasing DNA sequence identity in comparing T4 with RB69 (another T-even phage) (27), RB49 (a pseudo-T-even *E. coli* phage), 44RR2.8t (a pseudo-T-even *Aeromonas* phage), and Aeh1 (a schizo-T-even *Aeromonas* phage). This relationship became even closer when comparing T4-like phages belonging to the same branch (e.g., the pseudo-T-even *E. coli* phages RB49 and  $\phi$ 1; the *Aeromonas* phages 44RR2.8t, 25, and 31; and the schizo-T-even *Aeromonas* phages Aeh1 and 65). However, none of these analyses have been published in greater detail. The focus of our analysis is comparisons between phages belonging to different branches of the T-even group of *E. coli* phages.

Some genetic exchanges between T4-like phages were observed and can be selected in the laboratory (1), but they do not resemble lambda-type modular exchanges. An instructive case is provided by the tail fiber genes. T4 and RB69, despite being closely related in the structural genes, differ in their distal tail fibers, a common mechanism of host range extensions (35). An exchange point apparently exists between proximal and distal tail fibers. In contrast, JS98 and RB69 share related distal tail fiber genes, while the proximal tail fiber genes differ substantially in sequence. Interestingly, this difference comes in parallel with differences in base plate wedge and base plate hub genes that likely interact with the proximal tail fiber genes. T4 phages can thus make coordinated exchanges over three separate genome regions when the proteins are adjacent in the virion structure and therefore have to interact directly.

Despite this flexibility, T4 phages display clear constraints in the choice of genes fulfilling similar functions. While genes lacking sequence relatedness can serve the same structural functions in lambdoid phages, T-even phages have to use genes derived from the same sequence family. Far fewer constraints exist for the left genome halves of T-even phages. A patchwork of conservation and diversity was revealed by the alignment of the phage genomes. Only a few genes shared high sequence identity across all T-even phages (e.g., the *g39* topoisomerase gene), most showed a variable degree of sequence conservation, and a few regions lacked any sequence relatedness. The alignment identified four such regions; two regions showed a group of genes which were unrelated in T4, RB69, and JS98, while two other regions showed distinct genes in two phages and no genes in the third phage. One of the hypervariable regions carries an intron endonuclease gene (*segA*) and two DNA modification genes; the other carries a tRNA gene cluster, again accompanied by an intron endonuclease gene (*segB*) (17, 25). Upstream of the soluble lysozyme gene *e*, T4 lacks genes where both JS98 and RB69 show a large cluster of unrelated genes. Where T4 carries the RNase reductase subunit gene *nrdD*, flanked on both sides by intron endonuclease genes, RB69 showed *nrdD* genes without endonuclease genes and JS98 showed no genes. Genetic hypervariability within T-even phages thus seems to be associated with mobile DNA. Since T4 has been invaded heavily by intron endonucleases, further phage comparisons are needed to assess whether this diagnosis can be generalized.

The T4 phage genome shows a peculiar strand-specific distribution of conserved and variable genes. The Watson strand carries the structural genes nearly exclusively and is highly conserved. The Crick strand carries the nonstructural genes. Over the left genome half, the Crick strand does not show a marked clustering of conserved gene function in adjacent genes. Notably, all JS98-specific genes are found exclusively on the Crick strand. The dispersed location of the JS98-specific genes within conserved phage gene functions (e.g., DNA replication) suggests that the strain-specific genes were inserted between the conserved nonstructural genes. In their majority, the inserted genes did not arrive as functional clusters but as individual genes or, at most, two adjacent genes. Novelty in T4-like phages comes primarily with these new genes and secondarily by the duplication of existing phage genes. In addition to the previously described *g24* duplication, we detected other adjacent JS98 genes with significant amino acid sequence identity (*alt*, *frd.2*, and *modB* genes). Since it was argued that the T4 gene *24* is already a duplication of gene *23* (16), duplication followed by sequence diversification might be a mode of T4 evolution, which became possible with the larger T4 genome.

Currently, we are comparing the genetic variability within the individual branches of the T-even phage group to gain further insight into the processes introducing genetic variability in T4 phages over even shorter periods.

#### ACKNOWLEDGMENTS

We thank Bernard Berger for providing initial information about the pyrosequencing technique at 454 Life Sciences and Francoise Delmas-Julien for her assistance with BioNumerics software. We thank Henry Krisch (CNRS Toulouse) for critically reading the manuscript and for helpful comments.

#### REFERENCES

1. Abe, M., Y. Izumoji, and Y. Tanji. 2007. Phenotypic transformation including host-range transition through superinfection of T-even phages. *FEMS Microbiol. Lett.* **269**:145–152.
2. Albert, M. J., A. S. Faruque, S. M. Faruque, R. B. Sack, and D. Mahalanabis. 1999. Case-control study of enteropathogens associated with childhood diarrhea in Dhaka, Bangladesh. *J. Clin. Microbiol.* **37**:3458–3464.
3. Bhan, M. K., D. Mahalanabis, O. Fontaine, and N. F. Pierce. 1994. Clinical trials of improved oral rehydration salt formulations: a review. *Bull. W. H. O.* **72**:945–955.
4. Black, R. E., M. H. Merson, A. S. Rahman, M. Yunus, A. R. Alim, I. Huq, R. H. Yolken, and G. T. Curlin. 1980. A two-year study of bacterial, viral, and parasitic agents associated with diarrhea in rural Bangladesh. *J. Infect. Dis.* **142**:660–664.
5. Brüssow, H. 2005. Phage therapy: the *Escherichia coli* experience. *Microbiology* **151**:2133–2140.
6. Brüssow, H., C. Canchaya, and W. D. Hardt. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68**:560–602.
7. Bruttin, A., and H. Brüssow. 2005. Human volunteers receiving *Escherichia coli* phage T4 orally: a safety test of phage therapy. *Antimicrob. Agents Chemother.* **49**:2874–2878.
8. Canchaya, C., C. Proux, G. Fournous, A. Bruttin, and H. Brüssow. 2003. Prophage genomics. *Microbiol. Mol. Biol. Rev.* **67**:238–276.
9. Chibani-Chennoufi, S., C. Canchaya, A. Bruttin, and H. Brüssow. 2004. Comparative genomics of the T4-like *Escherichia coli* phage JS98: implications for the evolution of T4 phages. *J. Bacteriol.* **186**:8276–8286.
10. Chibani-Chennoufi, S., J. Sidoti, A. Bruttin, M. L. Dillmann, E. Kutter, F. Qadri, S. A. Sarker, and H. Brüssow. 2004. Isolation of *Escherichia coli* bacteriophages from the stool of pediatric diarrhea patients in Bangladesh. *J. Bacteriol.* **186**:8287–8294.
11. Comeau, A. M., C. Bertrand, A. Letarov, F. Tétart, and H. M. Krisch. 2007. Modular architecture of the T4 phage superfamily: a conserved core genome and a plastic periphery. *Virology* **362**:384–396.
12. Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**:4636–4641.
13. Desplats, C., C. Dez, F. Tétart, H. Eleaume, and H. M. Krisch. 2002. Snapshot of the genome of the pseudo-T-even bacteriophage RB49. *J. Bacteriol.* **184**:2789–2804.
14. Dupont, H. L., Z. D. Jiang, J. Belkind-Gerson, P. C. Okhuysen, C. D. Ericsson, S. Ke, D. B. Huang, M. W. Dupont, J. A. Adachi, F. J. De La Cabada, D. N. Taylor, S. Jaini, and S. F. Martinez. 2007. Treatment of travelers' diarrhea: randomized trial comparing rifaximin, rifaximin plus loperamide, and loperamide alone. *Clin. Gastroenterol. Hepatol.* **5**:451–456.
15. Filée, J., F. Tétart, C. A. Suttle, and H. M. Krisch. 2005. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. USA* **102**:12471–12476.
16. Fokine, A., P. G. Leiman, M. M. Shneider, B. Ahvazi, K. M. Boeshans, A. C. Steven, L. W. Black, V. V. Mesyanzhinov, and M. G. Rossmann. 2005. Structural and functional similarities between the capsid proteins of bacteriophages T4 and HK97 point to a common ancestry. *Proc. Natl. Acad. Sci. USA* **102**:7163–7168.
- 16a. Golomidova, A., E. Kulikov, A. Isaera, A. Manykin, and A. Letarov. 2007. The diversity of coliphages and coliforms in horse feces reveals a complex pattern of ecological interactions. *Appl. Environ. Microbiol.* **73**:5975–5981.
17. Kadyrov, F. A., M. G. Shlyapnikov, and V. M. Kryukov. 1997. A phage T4 site-specific endonuclease, SegE, is responsible for a non-reciprocal genetic exchange between T-even-related phages. *FEBS Lett.* **415**:75–80.
18. Karam, J. D. (ed.). 1994. Molecular biology of bacteriophage T4. American Society for Microbiology, Washington, DC.
19. Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* **5**:R12.
20. 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.
21. Mann, N. H., M. R. Clokie, A. Millard, A. Cook, W. H. Wilson, P. J. Wheatley, A. Letarov, and H. M. Krisch. 2005. The genome of S-PM2, a "photosynthetic" T4-type bacteriophage that infects marine *Synechococcus* strains. *J. Bacteriol.* **187**:3188–3200.
22. Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P.

- Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376–380.
23. Merson, M. H., R. B. Sack, S. Islam, G. Saklayen, N. Huda, I. Huq, A. W. Zulich, R. H. Yolken, and A. Z. Kapikian. 1980. Disease due to enterotoxigenic *Escherichia coli* in Bangladeshi adults: clinical aspects and a controlled trial of tetracycline. *J. Infect. Dis.* **141**:702–711.
  24. Miller, E. S., J. F. Heidelberg, J. A. Eisen, W. C. Nelson, A. S. Durkin, A. Ciecko, T. V. Feldblyum, O. White, I. T. Paulsen, W. C. Nierman, J. Lee, B. Szczypinski, and C. M. Fraser. 2003. Complete genome sequence of the broad-host-range vibriophage KVP40: comparative genomics of a T4-related bacteriophage. *J. Bacteriol.* **185**:5220–5233.
  25. Miller, E. S., E. Kutter, G. Mosig, F. Arisaka, T. Kunisawa, and W. Ruger. 2003. Bacteriophage T4 genome. *Microbiol. Mol. Biol. Rev.* **67**:86–156.
  26. Myers, E., and W. Miller. 1988. Optimal alignments in linear space. *Comput. Appl. Biol. Sci.* **4**:11–17.
  27. Nolan, J. M., V. Petrov, C. Bertrand, H. M. Krisch, and J. D. Karam. 2006. Genetic diversity among five T4-like bacteriophages. *Virology* **3**:30.
  28. Petrov, V. M., J. M. Nolan, C. Bertrand, D. Levy, C. Desplats, H. M. Krisch, and J. D. Karam. 2006. Plasticity of the gene functions for DNA replication in the T4-like phages. *J. Mol. Biol.* **361**:46–68.
  29. Qadri, F., T. Ahmed, F. Ahmed, S. R. Bradley, D. A. Sack, and A. M. Svennerholm. 2003. Safety and immunogenicity of an oral, inactivated enterotoxigenic *Escherichia coli* plus cholera toxin B subunit vaccine in Bangladeshi children 18–36 months of age. *Vaccine* **21**:2394–2403.
  30. Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. 2000. Artemis: sequence visualization and annotation. *Bioinformatics* **16**:944–945.
  31. Savarino, S. J., E. R. Hall, S. Bassily, T. F. Wierzbza, F. G. Youssef, L. F. Peruski, Jr., R. Abu-Elyazeed, M. Rao, W. M. Francis, H. El Mohamady, M. Safwat, A. B. Naficy, A. M. Svennerholm, M. Jertborn, Y. J. Lee, and J. D. Clemens. 2002. Introductory evaluation of an oral, killed whole cell enterotoxigenic *Escherichia coli* plus cholera toxin B subunit vaccine in Egyptian infants. *Pediatr. Infect. Dis. J.* **21**:322–330.
  32. Snyder, J. D., and M. H. Merson. 1982. The magnitude of the global problem of acute diarrhoeal disease: a review of active surveillance data. *Bull. W. H. O.* **60**:605–613.
  33. Sullivan, M. B., M. L. Coleman, P. Weigele, F. Rohwer, and S. W. Chisholm. 2005. Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.* **3**:e144.
  34. Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. A genomic perspective on protein families. *Science* **278**:631–637.
  35. Tétart, 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.
  36. Tétart, F., C. Desplats, M. Kutateladze, C. Monod, H. W. Ackermann, and H. M. Krisch. 2001. Phylogeny of the major head and tail genes of the wide-ranging T4-type bacteriophages. *J. Bacteriol.* **183**:358–366.
  37. Warner, H. R., D. P. Snustad, J. F. Koerner, and J. D. Childs. 1972. Identification and genetic characterization of mutants of bacteriophage T4 defective in the ability to induce exonuclease A. *J. Virol.* **9**:399–407.
  38. Yeh, L. S., T. Hsu, and J. D. Karam. 1998. Divergence of a DNA replication gene cluster in the T4-related bacteriophage RB69. *J. Bacteriol.* **180**:2005–2013.