Running title: Phylogenetics of dU jumbophages

Phylogenic analysis of new viral cluster of large phages with unusual DNA genomes containing uracil in place of thymine in gene-sharing network, using phages S6 and PBS1 and relevant uncultured phages derived from sewage metagenomics

Jumpei Uchiyama[1*], Iyo Takemura-Uchiyama[1], Kazuyoshi Gotoh[1], Shin-ichiro Kato[2], Yoshihiko Sakaguchi[3], Hironobu Murakami[4], Tomoki Fukuyama[4], Mao Kaneki[4], Osamu Matsushita[1], Shigenobu Matsuzaki[5]

[1] Department of Bacteriology, Graduate School of Medicine Dentistry and Pharmaceutical Sciences, Okayama University, Okayama 700-8558, Japan.

[2] Research Institute of Molecular Genetics, Kochi University, Kochi 783-0093, Japan.

[3] Department of Microbiology, Kitasato University School of Medicine, Kanagawa 252-0374, Japan.

[4] School of Veterinary Medicine, Azabu University, Kanagawa 252-5201, Japan.

[5] Department of Medical Laboratory Science, Faculty of Health Sciences, Kochi Gakuen University, Kochi 780-0955, Japan.

* Corresponding author.

Mailing address: Shikata-cho 2-5-1, Kita-ku Okayama-shi, Okayama 700-8558, Japan.

Phone: +81-86-235-7158. E-mail: uchiyama@okayama-u.ac.jp.

ABSTRACT

Bacteriophages (phages) are the most diverse and abundant life-form on Earth. Jumbophages are phages with double-stranded DNA genomes longer than 200 kbp. Among these, some jumbophages with uracil in place of thymine as a nucleic acid base, which we have tentatively termed “dU jumbophages” in this study, have been reported. Because the dU jumbophages are considered to be a living fossil from the RNA world, the evolutionary traits of dU jumbophages are of interest. In this study, we examined the phylogeny of dU jumbophages. First, tBLASTx analysis of newly sequenced dU jumbophages such as *Bacillus* phage PBS1 and previously isolated *Staphylococcus* phage S6 showed similarity to the other dU jumbophages. Second, we detected the two partial genome sequences of uncultured phages possibly relevant to dU jumbophages, scaffold_002 and scaffold_007, from wastewater metagenomics. Third, according to the gene-sharing network analysis, the dU jumbophages, including phages PBS1 and S6, and uncultured phage scaffold_002 formed a cluster, which suggested a new viral subfamily/family. Finally, analyses of the phylogenetic relationship with other phages showed that the dU jumbophage cluster, which had two clades of phages infecting Gram-negative and Gram-positive bacteria, diverged from the single ancestral phage. These findings together with previous reports may imply that dU jumbophages evolved from the same origin before divergence of Gram-negative and Gram-positive bacteria.

Keywords: environmental virus; jumbophage; metagenomics; evolution; uncultured phage.

1. Introduction

In recent years, many bacteriophages (phages) have been isolated and studied because of the increasing interest in biology, ecology, and phage therapy; a number of novel uncultured phages have been discovered because of the development of high-throughput sequencer technologies and lowering costs. Along with the research, phages with relatively large genomes and genome sizes similar to small-sized bacteria, such as *Mycoplasma genitalium* and leafhopper symbiotic bacteria *Karelsulcia muelleri* and *Nasuia deltocephalinicola*, have been discovered (Bennett and Moran, 2013; Fraser et al., 1995; Hendrix, 2009; Turner et al., 2021). Regarding such large phages, those with genomes > 200 kb have been referred to as jumbophages (Yuan and Gao, 2017) and those with genomes > 500 kb as megaphages (Devoto et al., 2019).

The large phages have attracted interest because of various aspects. Such phages have unique biological features such as the life cycle of lysis and pseudolysogeny, a host takeover mechanism, and a prevention mechanism against superinfection (Al-Shayeb et al., 2020; Iyer et al., 2021). In addition to these features, comparative genomic research has suggested that large phages are considered to have evolved from smaller phages, and to have emerged in and before the period of the last universe of common ancestor (LUCA) (Iyer et al., 2021; Nazir et al., 2021). However, the study of large phages is limited compared with smaller phages, because of two technological difficulties for phage isolation and metagenomics.

First, the isolation of large phages with large genomes by the ordinal method of plaque assay remains difficult because of inefficient diffusion in the agarose gel and the different culture conditions from ordinary conditions of bacteria and phages (Serwer and Wright, 2020; Uchiyama et al., 2014). In addition, the diversity of isolated phages is limited because the host bacteria used to isolate phages are biased (Cook et al., 2021). Moreover, although metagenomic analysis has discovered a number of phages with large genomes from various sources such as the environment and humans (Al-Shayeb et al., 2020; Devoto et al., 2019; Hurwitz et al., 2018; Yahara et al., 2021), the specific DNA modification of phages cannot be read by the common metagenomic method (Rihtman et al., 2021). Thus, the study of phages with large genomes can be accelerated by both the classical method of phage isolation and the exploratory method of metagenomics.

Some phages including large ones appear to have various unique DNA modifications (Hutinet et al., 2021). Among these, some phages contain uracil instead of thymine as a nucleic acid base in their double-stranded DNA genome, which is tentatively termed "dU phages" in this study. The first dU phages discovered were a group of jumbophages such as *Bacillus* phages PBS1 and AR9, *Yersinia* phage phiR1-37, and *Staphylococcus* phage S6, which we have isolated (Hunter et al., 1967; Kiljunen et al., 2005; Lavysh et al., 2016; Uchiyama et al., 2014). In recent years, smaller dU phages have been discovered, which are globally distributed (Rihtman et al., 2021). Because dU jumbophages are considered to be a remnant from the RNA world (Nagy et al., 2021), phylogenetic analysis of them among a variety of phages may provide a clue to unraveling the mystery of the evolution of bacteria and large phages. However, the dU jumbophage group has hardly been characterized phylogenetically (Cook et al., 2021).

In this study, we phylogenetically characterized the dU jumbophages. First, we sequenced the genomes of previously isolated large dU jumbophages such as *Bacillus* phage PBS1 and previously isolated *Staphylococcus* phage S6. Second, we obtained the uncultured phage sequences relevant to the large dU jumbophage by metagenomic approach of sewage DNA. Third, we conducted gene-sharing network analysis among prokaryotic viruses. Finally, phylogenetic analyses based on large terminase and DNA polymerase.

2. Materials and Methods

2.1. Reagents and culture media

All reagents were purchased from Nacalai Tesque (Kyoto, Japan) or Fujifilm Wako Pure Chemicals (Osaka, Japan), unless otherwise stated. Luria-Bertani media (LB medium [Miller]; Kanto Chemical Co., Tokyo, Japan) was used.

2.2. Phage genome sequencing

Staphylococcus phage S6 has been isolated previously by us, as described elsewhere (Uchiyama et al., 2014). *Bacillus* phage PBS1 was obtained from Bacillus Genetic Stock Center, OH, USA (Takahashi, 1963). *S. aureus* strain SA27 and *B. subtilis* strain 168 were used as host bacteria for phages S6 and PBS1, respectively.

Phages were amplified at 30°C with appropriate host bacteria, and then purified by CsCl density-gradient centrifugation, as described elsewhere (Nasukawa et al., 2017). Genomic DNA was purified by the phenol-chloroform extraction method, as described elsewhere (Uchiyama et al., 2009). After multiple displacement amplification using GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Little Chalfont, United Kingdom), a shotgun library was prepared using the GS FLX Titanium rapid library preparation kit (Roche Diagnostics, Indianapolis, IN, USA) according to the manufacturer's instructions. The libraries were analyzed using a GS Junior 454 sequencer (Roche Diagnostics). The sequence reads were assembled using the 454 Newbler software (version 3.0; 454 Life Sciences, Branford, CT, USA) (sequence depth of S6 and PBS1 genome sequencing: 30 and 37, respectively). Based on the draft genome sequence, the genome sequence was proofread by the direct sequencing of both strands with a primer walking method using an ABI Prism 3100-Avant genetic analyzer (Applied Biosystems, Foster City, CA).

2.3. Sequencing of DNA obtained from sewage water

After debris removal by centrifugation from 250 mL of sewage influent water, polyethylene glycol 6000 and NaCl was supplemented at 10% and 0.5 M, respectively. After centrifugation (10,000 × g, 40 min, 4°C), the pellet suspended in 5 mL TM buffer (10 mM Tris-HCl [pH 7.2], 5 mM $MgCl_2$) was treated with 50 μg/mL DNase A and RNase I (30 min, 37°C). After centrifugation (10,000 × g, 3 min, 4°C), the supernatant was collected.

Mixing the supernatant sample with an equal amount of 2% (wt/vol) low-melting-temperature agarose, the plug was prepared. The plug was treated in a lysis solution (100 μg/mL protease K, 1% SDS, 10 mM Tris–HCl pH 8.0, 1 mM EDTA) for 24 h at 50°C. The plug was washed with TBE buffer twice. The DNA was separated using a CHEF Mapper apparatus (Bio-Rad Laboratories, Hercules, CA, USA) through a 1% (wt/vol) agarose gel (SeaKem Gold; FMC Bioproducts) in 0.5 × TBE buffer, together with a size marker (CHEF DNA Size Standard Lambda Ladder, Bio-Rad Laboratories). Switch times were ramped from 1 to 26 s over 22 h at 14°C and 6 V/cm.

The gel stained by ethidium bromide was visualized (Supplementary Fig. S1), and the gel was excised. The DNA was extracted from the gel using QIAquick Gel Extraction Kit (Qiagen, Venlo, Netherlands), and was amplified using GenomiPhi V2 DNA Amplification Kit (GE Healthcare). The library was prepared using Illumina TruSeq PCR-free DNA Library Preparation Kit, and was sequenced using Illumina HiSeq 2500 paired-end technology. The sequencing run yielded 23,260,816 filtered reads with 101-bp paired-end sequencing. The sequence data was trimmed using Trimmomatic v.0.32, and 23,112,810 reads were obtained (Bolger et al., 2014). The viral taxonomic classification was done using Kaiju v.1.8.2, the trimmed sequence was analyzed with greedy run mode at default setting against virus data from the NCBI RefSeq database (downloaded on February 5, 2022) (Menzel et al., 2016). The trimmed sequences were assembled using IDBA-UD v.1.1.1 (Peng et al., 2012).

2.4. Processing and analysis of sequence data

The genomes of phages belonging to the *Caudovirales* family, which are listed on the Genome Table, National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/genome/browse/#!/viruses/13352/), were selected by > 200 kbp in size. Three-hundred seventy-one genomes were downloaded from the GenBank database (accessed on January 29, 2022) and were used as a local database.

The sequence annotation was made using Dfast v.1.4.0 (https://dfast.ddbj.nig.ac.jp/) (Tanizawa et al., 2018). The protein function was predicted by MetaGeneAnnotator implemented in Dfast v.1.4.0 and InterProScan 5.54-87.0 (Blum et al., 2021). Sequences were analyzed using the BLAST program at the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and were locally analyzed using BLAST+ 2.6.0. (Altschul et al., 1997). The sequence processing, such as random sampling, size filtration, and sequence statistics, was done using SeqKit v.2.1.0 (Shen et al., 2016). The genome completeness was estimated by CheckV v0.9.0 using CheckV database v1.2 (Nayfach et al., 2021). The orthologous genes were predicted with default setting (BLASTp threshold score, 75) using Coregenes3.5 (http://binf.gmu.edu:8080/CoreGenes3.5/) (Zafar et al., 2002). The genome comparison by tBLASTx was visualized using the Easyfig v.2.2.2 (Sullivan et al., 2011). Host bacteria were predicted using VirHostMatcher-Net (downloaded on February 8, 2022) (Wang et al., 2020).

2.5. Gene-sharing network analysis

Viral proteins were analyzed by a network analysis of shared genes using vConTACT2 0.9.22 (with arguments "–rel-mode 'Diamond' –pcs-mode MCL –vcs-mode ClusterONE") against its ProkaryoticViralRefSeq211-Merged database (Bin Jang et al., 2019; Turner et al., 2021). The resulting network was visualized using Cytoscape 3.9.1 (Bin Jang et al., 2019; Shannon et al., 2003; Turner et al., 2021).

2.6. Phylogenetic analysis

The protein sequences of putative large terminase or DNA polymerase were subjected to delta-BLAST analysis to RefSeq protein database restricted by *Caudovirales* (taxonomy ID:28883), and all the object protein sequences were downloaded (accessed on 15 June, 2022). As query protein sequences of large terminase and DNA polymerase, gp014 and gp079 of *Staphylococcus* phage S6 (i.e., accession Nos., BDE75552 and BDE75617), gp114 and gp104 of *Bacillus* phage PBS1 (i.e., accession Nos., BDE75349 and BDE75339), and ORF081 and ORF096 of uncultured phage scaffold_002 (i.e., accession Nos., BDH16440 and BDH16455) were used, respectively. After the downloaded data were merged, duplicated protein sequences were removed, and protein sequences with lengths of 100–1,000 amino acids were extracted, resulting in 462 and 433 protein sequences for large terminase and DNA polymerase, respectively. One hundred protein sequences were randomly selected from each dataset, and the relevant protein sequences of dU jumbophages were merged. Data manipulations, such as duplicate removal, data extraction, and random sampling, were done using SeqKit v.2.1.0 (Shen et al., 2016).

The sequence alignment was done using ClustalW ver2.1 (Larkin et al., 2007), and the aligned sequences were trimmed using TrimAl v1.4.rev15 with the option '-automated1' (Capella-Gutierrez et al., 2009). The phylogenetic tree was constructed by the maximum likelihood method using IQ-TREE version 2.2.03 with 1,000 replicates for ultrafast bootstrap with model selection option (-m MFP) (Minh et al., 2020). The tree was visualized and manipulated using MEGA v.11.0.8. (Tamura et al., 2021).

2.7. Sequence data registry

The genome sequences of phages S6 and PBS1 were deposited to GenBank (accession Nos. LC680885 and LC680884, respectively).

The raw reads of wastewater metagenomes were deposited in the DNA Data

Bank of Japan (DDBJ) Read Archive (accession No. DRA013444). The two uncultured phage sequences, scaffold_002 and scaffold_007, which were derived from metagenomic data, were deposited to GenBank (accession Nos. LC701594 and LC701595, respectively).

3. Results and Discussion

3.1. Genome sequences of *Staphylococcus* phage S6 and *Bacillus* phage PBS1

The whole genomes of the previously isolated dU jumbophages, such as *Staphylococcus* phage S6 and *Bacillus* phage PBS1, were sequenced. Because DNA extracted from phage particles could not be read by 454 sequencing technology directly, the DNA amplified by multiple displacement amplification was sequenced by the 454 technology. The draft genomes, which contain the homopolymer produced by 454 sequencing technology, was then proofread by Sanger sequencing. The sequencing of complete genomes of phages S6 and PBS1 confirmed 267,055 bp and 252,136 bp, respectively. According to the annotation of the genomes of phages S6 and PBS1, 272 and 303 coding sequences (CDSs), and one tRNA gene were predicted in both phages. These phages had similar CDS numbers and G+C content to other dU jumbophages (Table 1).

We then analyzed the genome sequences of phages S6 and PBS1 by online BLASTn at the NCBI, and by local tBLASTx against 371 large phage genomes downloaded from the NCBI (Supplementary Table S1). First, according to the online BLASTn (January 20, 2022), phage S6 showed high genome-wide similarity to the other jumbophages, such as *Staphylococcus* phages PALS_2, vB_SauM-UFV_DC4, Madawaska, MarsHill, vB_StaM_SA1, and Machias (85.0%–98.6% in terms of identity and 63%–96% in terms of query coverage; Supplementary Table S2). Phage PBS1 showed high similarity to the other jumbophages, such as *Bacillus* phages AR9 and vB_BspM_Internexus (98% and 93% of query coverage and 99.6% and 96.9% of identity, respectively; Supplementary Table S2). Moreover, the tBLASTx of phages S6 and PBS1 to the local database detected 16 and 20 phages with a score of > 100, respectively, which included the other dU jumbophages such as AR9 and phiR1-37 (Supplementary Table S3).

3.2. Detection of genome sequences of uncultured phages relevant to dU

jumbophages from metagenomic data derived from size-selected sewage water DNA

We searched for the uncultured phages relevant to dU jumbophages from wastewater using a metagenomic approach in this study. After removal of bacteria and debris by centrifugation, the DNA was separated by pulsed-field gel electrophoresis. The gel located at ca. 200–340 kbp was excised, and DNA was purified. The DNA was amplified by multiple displacement amplification, and the sequencing was performed. The 2.3-Gb short-read sequence data were obtained and trimmed for the following analysis. To observe an overview of the sequence data, the trimmed sequences were taxonomically assigned to the RefSeq database, using the metagenomic pipeline. Ca. 8.0% of reads were shown to be assigned as viruses. When the total ratio of virus-assigned sequences was set at 100%, the order *Caudovirales* occupied 95%, and the rest was other viral taxa, at 5% (Supplementary Fig. S2).

We then searched for the genome sequences relevant to dU jumbophages from the sequence data. First, the assembly of trimmed reads produced 88,325 scaffold sequences. Because the scaffold data contained many short sequences, scaffold sequences less than 20 kb were removed. As a result, 200 scaffold sequences were obtained (total 6,370,609 nt, and mean 31,853 nt in length). Subsequently, 200 scaffold sequences were filtered by similarity to dU jumbophages using tBLASTx. The sequences of dU jumbophages such as S6, PBS1, AR9, and phiR1-37 were compared with 200 scaffold sequences with a cutoff e-value of 1E-04. As a result, 63, 82, 85, and 79 scaffold sequences were detected for phages S6, AR9, PBS1, and phiR1-37, respectively (Supplementary Table S4), of which 40 scaffold sequences were detected in common.

Among these 40 common scaffold sequences, two large sequences, 200,670 bp and 111,289 bp, were present (scaffold_002 and scaffold_007, respectively); the 38 other ranged from 20 kb to 55 kb (Table 2; Supplementary Table S5). Estimating the completeness and complete genome size these 40 scaffolds by the CheckV program, 33 were assumed to be partial genome with larger than 200 kbp. In particular, scaffold_002 and scaffold_007 were estimated to be 262,391 bp and 272,963 bp in length, the genome completeness of which was predicted to be 76.5% and 40.8%, respectively (Supplementary Table S5).

These two large scaffolds, scaffold_002 and scaffold_007, were characterized by comparison with other large phages. The scaffold sequences were analyzed by local BLAST to large phage sequences. First, the local BLASTn analysis of scaffold_002 and scaffold_007 showed no sequences with high coverage. Subsequently, scaffold_002 and scaffold_007 were analyzed by local tBLASTx to large phage sequences. The scaffold_002 sequence analysis showed *Yersinia* phage phiR1-37 (score 797, E-value 0) as a top hit (Table 2) and *Bacillus* phage AR9 as the fourth-highest hit (Supplementary Table S6); the scaffold_007 sequence analysis showed that dU jumbophage was not detected among the top 10 hits, while it showed that large non-dU phage *Ralstonia* phage RP31 (score 232, E-value 0) was detected as a top hit (Supplementary Table S6).

Annotating the scaffold_002 and scaffold_007 sequences, 211 CDSs with one tRNA gene and 111 CDSs were predicted, respectively (Supplementary Tables S7 and S8). The orthologous genes between scaffold_002 and *Yersinia* phage phiR1-37 and between scaffold_007 and *Ralstonia* phage RP31 were analyzed by Coregenes software. Scaffold_007 was predicted to be 48.8% (103/211 CDSs) orthologous to CDS of *Yersinia* phage phiR1-37. Scaffold_002 was predicted to be 27.0% (30/111 CDSs) orthologous to CDS of *Ralstonia* phage RP31. Moreover, observing the arrangement of orthologous genes (Supplementary Tables S7 and S8) in the scaffold_002 and scaffold_007 sequences, the gene arrangement of scaffold_002 was synchronized to *Yersinia* phage phiR1-37. In addition, examining the genome-wide similarity by tBLASTx, genome-wide synteny was observed between scaffold_002 and *Yersinia* phage phiR1-37, while synteny was partially observed on scaffold_007 to *Ralstonia* phage RP31 (Fig. 1). Considering these results, two large scaffolds, scaffold_002 and scaffold_007, were considered to be partial genomes of uncultured jumbophages possibly relevant to dU jumbophages.

In recent years, host prediction tools have been developed, and they can be categorized into three main types: alignment-dependent, alignment-independent, and integrative methods (Coclet and Roux, 2021). The integrative method is presently the most promising method, whereby a combination of several methods leads to a single prediction. We attempted to predict the host bacteria of uncultured jumbophages scaffold_002 and scaffold_007 using the integrative method. The host bacteria of scaffold_002 and scaffold_007 were predicted to be genera *Staphylococcus* and *Acinetobacter*, respectively (Table 2).

3.3. Taxonomic assignment of dU jumbophages and uncultured jumbophages by gene-sharing network analysis

To characterize the dU jumbophages and uncultured jumbophages from a taxonomical point of view, we conducted the taxonomic assignment of *Staphylococcus* phage S6 and *Bacillus* phage PBS1, and uncultured phages scaffold_002 and scaffold_007 using the gene-sharing network analysis tool vContact2 against all the phages (i.e., small to large phages) in the RefSeq (Bin Jang et al., 2019; Turner et al., 2021). As a result, *Staphylococcus* phage S6 and *Bacillus* phage PBS1, and uncultured phage scaffold_002 were located in the same viral cluster as phages *Yersinia* phage phiR1-37 and *Bacillus* phage AR9, which was located at the end on the largest network containing multiple viral clusters and was branched off from the cluster of subfamily *Twortvirinae*.

On the other hand, uncultured phage scaffold_007 was located on the independent viral cluster containing other 43 phages, which suggested the other viral subfamily/family of large phages. In the viral cluster, scaffold_007 had links to 37 jumbophages, including *Pseudomonas* phage 201phi2-1, *Serratia* phage Moabite, *Ralstonia* phage RP12, *Ralstonia* phage RSL2, *Pseudomonas* phage phiKZ, *Erwinia* phage phiEaH1, and *Escherichia* phage vB_EcoM_Goslar. Thus, apart from the dU jumbophages, uncultured phage scaffold_007 seemed to be a different type of jumbophage, for which the viral cluster has not been designated taxonomically to date.

Considering these results, *Bacillus* phage PBS1, *Bacillus* phage AR9, *Staphylococcus* phage S6, *Yersinia* phage phiR1-37, and uncultured phage scaffold_002, can be grouped as a new viral subfamily/family of dU jumbophages.

3.4. Phylogenetic analysis of dU jumbophages based on large terminase and DNA polymerase

Large terminase can be used for phylogenetic analysis for large phages (Al-Shayeb et al., 2020), and DNA polymerase also can be used for phylogenetic analysis among large phages including dU jumbophages (Iyer et al., 2021). We constructed the phylogenetic trees based on large terminase and DNA polymerase, and analyzed the phylogenetic relationship of dU jumbophages with other relevant phage proteins. First, according to the phylogenetic tree based on large terminase (Fig. 3A), the dU jumbophages were clustered in the tree, as with the gene-sharing network. In the tree, *Bacillus* phages AR9 and PBS1 were branched off from the same node as *Yersinia* phage phiR1-37 and uncultured phage scaffold_002; *Staphylococcus* phage was located separately from these dU jumbophages. Other jumbophages were also clustered and were sparsely located in the tree. Several non-jumbophages were observed among jumbophages.

Next, according to the phylogenetic tree based on DNA polymerase, the dU jumbophages were clustered, similar to the phylogenetic tree based on large terminase. In the tree, *Bacillus* phages AR9 and PBS1 together with *Staphylococcus* phage S6 were branched off from the same node as *Yersinia* phage phiR1-37 and uncultured phage scaffold_002. Although the other jumbophages were present next to the dU jumbophage cluster, *Vibrio* phage JM-2012 was present in the middle. Although *Vibrio* phage JM-2012 is ca 167 kbp in genome size, it is considered to be related to *Pseudomonas* jumbophage phiKZ (Jang et al., 2013). *Vibrio* phage JM-2012 has no protein sequence similarity to the DNA polymerase of globally-distributed smaller dU phages (*i.e.*, roseophages DSS3_VP1 and DSS3_PM1) (Rihtman et al., 2021).

In both trees, the dU jumbophages were clustered, apart from the other jumbophages, suggesting that they originated from a common ancestral phage. Jumbophages are considered to originate from several smaller phages through multiple processes, and small dU phages have been discovered (Iyer et al., 2021; Rihtman et al., 2021). Thus, because dU jumbophages were considered to be one type of jumbophages based on our result, the dU jumbophages may originate from the same ancestral smaller phage.

3.5. Evolutional implication of dU jumbophages

When considering the phage evolution of a specific phage linage, host bacteria can be used as a predictor for phage evolution. In this host prediction, we also used a VirHostMatcher-Net software as a host prediction tool, which is believed to have one of the best prediction reliabilities to date (Coclet and Roux, 2021). Because the accuracy of this software is not certain among large phages, the prediction accuracy was examined. Examining the correct matches between the actual host and the predicted host from 302 large phages, the correct match rates at levels of phylum, class, order, family, and genus levels were 63.9%, 59.9%, 48.0%, 14.9%, and 12.3%, respectively (Supplementary Fig. S3). These match rates were not as high as expected. This is probably because of insufficient phage-host information of large phages in the database.

Although the host bacteria of uncultured phage scaffold_002 was predicted to be the genus *Staphylococus* spp., the assumption from the phylogenetic trees suggested that the host bacteria of scaffold_002 was Gram-negative bacteria. We tentatively assumed that the host bacteria of uncultured phage scaffold_002 was Gram-negative bacteria, because of implication from the phylogenetic analysis, and the dU jumbophages infecting Gram-negative and Gram-positive bacteria appeared to diverge at some time point in the past. Monoderms and diderms were divergently evolved from ancestor cells in the bacterial evolution (Megrian et al., 2020). Considering these, the dU jumbophages may emerge from the same ancestral smaller predecessor before divergence of Gram-positive and Gram-negative bacteria. We believe that this evidence also supports the emergence of large phages at or before the period of the LUCA.

One of the strengths of our study is the combination of the phage isolation and metagenomics approaches. Phage isolation remains a very powerful experimental approach, as the discovery of novel phages can produce a large amount of basic information; and metagenomic analysis allows for the efficient search of uncultured phages. We believe that such an approach will enhance the accumulation of knowledge for the dU jumbophage group and contribute to the elucidation of bacterial and phage evolution.

Funding

This project was in funded by Grant-in-Aid for Young Scientists (B) (Grant No. 15K19095) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and the Sasakawa Scientific Research Grant from The Japan Science Society (Grant No. 28-604).

Author contributions

J.U.: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration, Funding acquisition. I.T-U.: Validation, Investigation, Funding acquisition. K.G.: Validation, Formal analysis. S.K.: Investigation. Y.S.: Investigation. H.M.: Writing - Review & Editing. T.F.: Writing -

Review & Editing. M.K.: Data Curation. O.M.: Writing - Review & Editing. S.M.: Investigation, Writing - Review & Editing.

Ethical statement

The authors declare no ethical issues relevant to this work.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest. We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work.

References

Al-Shayeb, B., Sachdeva, R., Chen, L.X., Ward, F., Munk, P., Devoto, A., Castelle, C.J., Olm, M.R., Bouma-Gregson, K., Amano, Y., He, C., Meheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D.S.A., Borton, M.A., Sharrar, A., Jaffe, A.L., Nelson, T.C., Kantor, R., Keren, R., Lane, K.R., Farag, I.F., Lei, S., Finstad, K., Amundson, R., Anantharaman, K., Zhou, J., Probst, A.J., Power, M.E., Tringe, S.G., Li, W.J., Wrighton, K., Harrison, S., Morowitz, M., Relman, D.A., Doudna, J.A., Lehours, A.C., Warren, L., Cate, J.H.D., Santini, J.M., Banfield, J.F., 2020. Clades of huge phages from across Earth's ecosystems. Nature 578(7795), 425-431.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17), 3389-3402.

Bennett, G.M., Moran, N.A., 2013. Small, smaller, smallest: the origins and evolution of ancient dual symbioses in a Phloem-feeding insect. Genome Biol Evol 5(9), 1675-1688.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., Turner, D., Sullivan, M.B., 2019. Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat Biotechnol 37(6), 632-639.

Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G.A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D.H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D.A., Necci, M., Orengo, C.A., Pandurangan, A.P., Rivoire, C., Sigrist, C.J.A., Sillitoe, I., Thanki, N., Thomas, P.D., Tosatto, S.C.E., Wu, C.H., Bateman, A., Finn, R.D., 2021. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49(D1), D344-D354.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15), 2114-2120.

Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15), 1972-1973.

Coclet, C., Roux, S., 2021. Global overview and major challenges of host prediction methods for uncultivated phages. Curr Opin Virol 49, 117-126.

Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D., Hobman, J., Jones, M.A., Millard, A., 2021. INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. PHAGE: Therapy, Applications, and Research 2(4), 214-223.

Devoto, A.E., Santini, J.M., Olm, M.R., Anantharaman, K., Munk, P., Tung, J., Archie, E.A., Turnbaugh, P.J., Seed, K.D., Blekhman, R., Aarestrup, F.M., Thomas, B.C., Banfield, J.F., 2019. Megaphages infect *Prevotella* and variants are widespread in gut microbiomes. Nat Microbiol 4(4), 693-700.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, R.D., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.F., Dougherty, B.A., Bott, K.F., Hu, P.C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., 3rd, Venter, J.C., 1995. The minimal gene complement of *Mycoplasma genitalium*. Science 270(5235), 397-403.

Hendrix, R.W., 2009. Jumbo bacteriophages. Curr Top Microbiol Immunol 328, 229-240.

Hunter, B.I., Yamagishi, H., Takahashi, I., 1967. Molecular weight of bacteriophage PBS 1 deoxyribonucleic acid. J Virol 1(4), 841-842.

Hurwitz, B.L., Ponsero, A., Thornton, J., Jr., U'Ren, J.M., 2018. Phage hunters:

computational strategies for finding phages in large-scale 'omics datasets. Virus Res 244, 110-115.

Hutinet, G., Lee, Y.J., de Crecy-Lagard, V., Weigele, P.R., 2021. Hypermodified DNA in viruses of *E. coli* and Salmonella. EcoSal Plus 9(2), eESP00282019.

Iyer, L.M., Anantharaman, V., Krishnan, A., Burroughs, A.M., Aravind, L., 2021. Jumbo phages: a comparative genomic overview of core functions and adaptions for biological conflicts. Viruses 13(1), 63.

Jang, H.B., Fagutao, F.F., Nho, S.W., Park, S.B., Cha, I.S., Yu, J.E., Lee, J.S., Im, S.P., Aoki, T., Jung, T.S., 2013. Phylogenomic network and comparative genomics reveal a diverged member of the PhiKZ-related group, marine vibrio phage PhiJM-2012. J Virol 87(23), 12866-12878.

Kiljunen, S., Hakala, K., Pinta, E., Huttunen, S., Pluta, P., Gador, A., Lonnberg, H., Skurnik, M., 2005. Yersiniophage phiR1-37 is a tailed bacteriophage having a 270 kb DNA genome with thymidine replaced by deoxyuridine. Microbiology (Reading) 151(Pt 12), 4093-4102.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21), 2947-2948.

Lavysh, D., Sokolova, M., Minakhin, L., Yakunina, M., Artamonova, T., Kozyavkin, S., Makarova, K.S., Koonin, E.V., Severinov, K., 2016. The genome of AR9, a giant transducing *Bacillus* phage encoding two multisubunit RNA polymerases. Virology 495, 185-196.

Megrian, D., Taib, N., Witwinowski, J., Beloin, C., Gribaldo, S., 2020. One or two membranes? diderm *Firmicutes* challenge the Gram-positive/Gram-negative divide. Mol Microbiol 113(3), 659-671.

Menzel, P., Ng, K.L., Krogh, A., 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 7, 11257.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37(5), 1530-1534.

Nagy, K.K., Skurnik, M., Vertessy, B.G., 2021. Viruses with U-DNA: New Avenues for Biotechnology. Viruses 13(5), 875.

Nasukawa, T., Uchiyama, J., Taharaguchi, S., Ota, S., Ujihara, T., Matsuzaki, S., Murakami, H., Mizukami, K., Sakaguchi, M., 2017. Virus purification by CsCl density gradient using general centrifugation. Arch Virol 162(11), 3523-3528.

Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., Kyrpides, N.C., 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat Biotechnol 39(5), 578-585.

Nazir, A., Ali, A., Qing, H., Tong, Y., 2021. Emerging aspects of jumbo bacteriophages. Infect Drug Resist 14, 5041-5055.

Peng, Y., Leung, H.C., Yiu, S.M., Chin, F.Y., 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11), 1420-1428.

Rihtman, B., Puxty, R.J., Hapeshi, A., Lee, Y.J., Zhan, Y., Michniewski, S., Waterfield, N.R., Chen, F., Weigele, P., Millard, A.D., Scanlan, D.J., Chen, Y., 2021. A new family of globally distributed lytic roseophages with unusual deoxythymidine to deoxyuridine substitution. Curr Biol 31(14), 3199-3206 e3194.

Serwer, P., Wright, E.T., 2020. In-gel isolation and characterization of large (and other) phages. Viruses 12(4), 410.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T., 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11), 2498-2504.

Shen, W., Le, S., Li, Y., Hu, F., 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One 11(10), e0163962.

Sullivan, M.J., Petty, N.K., Beatson, S.A., 2011. Easyfig: a genome comparison visualizer. Bioinformatics 27(7), 1009-1010.

Takahashi, I., 1963. Transducing phages for *Bacillus subtilis*. J Gen Microbiol 31, 211-217.

Tamura, K., Stecher, G., Kumar, S., 2021. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol 38(7), 3022-3027.

Tanizawa, Y., Fujisawa, T., Nakamura, Y., 2018. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 34(6), 1037-1039.

Turner, D., Kropinski, A.M., Adriaenssens, E.M., 2021. A roadmap for genome-based phage taxonomy. Viruses 13(3), 506.

Uchiyama, J., Rashel, M., Matsumoto, T., Sumiyama, Y., Wakiguchi, H., Matsuzaki, S., 2009. Characteristics of a novel *Pseudomonas aeruginosa* bacteriophage, PAJU2, which is genetically related to bacteriophage D3. Virus Res 139(1), 131-134.

Uchiyama, J., Takemura-Uchiyama, I., Sakaguchi, Y., Gamoh, K., Kato, S., Daibata, M., Ujihara, T., Misawa, N., Matsuzaki, S., 2014. Intragenus generalized transduction in *Staphylococcus* spp. by a novel giant phage. ISME J 8(9), 1949-1952.

Wang, W., Ren, J., Tang, K., Dart, E., Ignacio-Espinoza, J.C., Fuhrman, J.A., Braun, J., Sun, F., Ahlgren, N.A., 2020. A network-based integrated framework for predicting virus-prokaryote interactions. NAR Genom Bioinform 2(2), lqaa044.

Yahara, K., Suzuki, M., Hirabayashi, A., Suda, W., Hattori, M., Suzuki, Y., Okazaki, Y., 2021. Long-read metagenomics using PromethION uncovers oral bacteriophages and their interaction with host bacteria. Nat Commun 12(1), 27.

Yuan, Y., Gao, M., 2017. Jumbo bacteriophages: an overview. Front Microbiol 8, 403.

Zafar, N., Mazumder, R., Seto, D., 2002. CoreGenes: a computational tool for identifying and cataloging "core" genes in a set of small genomes. BMC Bioinformatics 3, 12.

Figure legends

Fig. 1. Comparison of scaffolds obtained from the sewage metagenomics with relevant phage genomes. The analyzed data by tBLASTx was visualized. Comparison of (A) scaffold_002 with *Yersinia* phage phiR1-37, and (B) scaffold_007 with *Ralstonia* phage RP31. The BLAST identity is shown as a scale bar at the bottom of each genome comparison figure. The genome size scale bar is shown below each genome comparison figure.

Fig. 2. Protein-sharing network of *Staphylococcus* phage S6, *Bacillus* phage PBS1, and uncultured phages scaffold_002 and scaffold_007 with prokaryotic virus data derived from RefSeq211. (A) Location of viral clusters containing the analyzed phage sequences. Each node and each edge between nodes represent a phage and phage connection based on pairwise shared protein content, respectively. The viral clusters containing analyzed phages (i.e., clusters A and B) were circled in red. Yellow nodes represent the analyzed phages in this study. (B) Viral cluster A. (C) Viral cluster B.

Fig. 3. Phylogenetic trees based on (A) large terminase and (B) DNA polymerase. Red dots represent the sequenced phages in this study. Phage names in red and blue are phages classified as dU jumbophages and other jumbophages, respectively.