UPΦ phages, a new group of filamentous phages found in several members of Enterobacteriales

Jason Shapiro
Loyola University Chicago, jshapiro2@luc.edu

Catherine Putonti
Loyola University Chicago, cputonti@luc.edu

Follow this and additional works at: https://ecommons.luc.edu/bioinformatics_facpub

Recommended Citation
Shapiro, Jason and Putonti, Catherine. UPΦ phages, a new group of filamentous phages found in several members of Enterobacteriales. Virus Evolution, 6, 1:, 2020. Retrieved from Loyola eCommons, Bioinformatics Faculty Publications, http://dx.doi.org/10.1093/ve/veaa030

This Article is brought to you for free and open access by the Faculty Publications and Other Works by Department at Loyola eCommons. It has been accepted for inclusion in Bioinformatics Faculty Publications by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution 4.0 International License.
© The Authors, 2020.
UPΦ phages, a new group of filamentous phages found in several members of Enterobacteriales

Jason W. Shapiro1,*,† and Catherine Putonti1,2,3,4

1Department of Biology, Loyola University Chicago, 1032 W Sheridan Rd, Chicago, IL 60660, USA, 2Department of Computer Science, Loyola University Chicago, 1052 W Loyola Ave, Chicago, IL, 60626, USA, 3Bioinformatics Program, Loyola University Chicago, 1052 W Loyola Ave, Chicago, IL 60626, USA and 4Department of Microbiology and Immunology, Stritch School of Medicine, Loyola University Chicago, 2160 S First Ave, Maywood, IL 60153, USA

*Corresponding author: E-mail: jshapiro2@luc.edu
†https://orcid.org/0000-0002-1321-8418

Abstract

Filamentous phages establish chronic infections in their bacterial hosts, and new phages are secreted by infected bacteria for multiple generations, typically without causing host death. Often, these viruses integrate in their host’s genome by co-opting the host’s XerCD recombinase system. In several cases, these viruses also encode genes that increase bacterial virulence in plants and animals. Here, we describe a new filamentous phage, UPΦ901, which we originally found integrated in a clinical isolate of Escherichia coli from urine. UPΦ901 and closely related phages can be found in published genomes of over 200 other bacteria, including strains of Citrobacter koseri, Salmonella enterica, Yersinia enterocolitica, and Klebsiella pneumoniae. Its closest relatives are consistently found in urine or in the blood and feces of patients with urinary tract infections. More distant relatives can be found in isolates from other environments, including sewage, water, soil, and contaminated food. Each of these phages, which we collectively call ‘UPΦ viruses’, also harbors two or more novel genes of unknown function.

Key words: bacteriophage; inovirus; prophage; bladder.

1. Introduction

Phages are often described by their potential to kill their hosts. Obligately lytic phages kill their hosts following infection, whereas temperate phages may lie dormant as prophages within lysogenized bacteria for several generations before entering a lytic cycle. While killing by phages has immediate consequences for bacterial ecology and has led to the revival of phage therapy for treating bacterial infections (Kortright et al. 2019), many phages also carry genes that alter bacterial behavior (Mai-Prochnow et al. 2015; Warwick-Dugdale et al. 2019). The effects of these phage-encoded genes range from modifying photosynthesis in cyanobacteria (Sieradzki et al. 2019) to producing toxins in potential pathogens (Waldor and Mekalanos 1996).
Inoviridae. Like many filamentous phages, UP Lineavirus urin sample. UP E. coli covered as a prophage in a clinical tives of UP E. coli Citrobacter koseri a tandem repeat at the dif locus in UMB0901 and only thirty-three with GenBank, and potato blight (Kamiunten and Wakimoto 1982; Yamada 2013). Across these various infections, filamentous phages may carry toxin genes, alter bacterial motility (jian, xiao, and wang 2013) and biofilm formation (rice et al. 2009; May, Tsuruta, and okabe 2011), or indirectly interfere with the immune system’s ability to clear the bacterial infection (sweere et al. 2019). Despite their myriad effects on a wide range of clinically and agriculturally important hosts, these phages remain understudied. As of this writing, there are only forty-six recognized members of Inoviridae in RefSeq, and only thirty-three with established taxonomy by the international committee on taxonomy of viruses (ICTV). Nonetheless, recent work has shown that thousands of potential inoviruses can be found in prophages in published bacterial genomes and metagenomes (Roux et al. 2019).

Here, we describe a novel filamentous phage, UPφ901, discovered as a prophage in a clinical E. coli isolate from a patient urine sample. UPφ901 is most closely related to the non-integrating phages I2-2 and IKE in the Lineavirus genus of the Inoviridae. Like many filamentous phages, UPφ901 integrates as a tandem repeat at the dif site in its hosts. Using the phage sequence from our isolate as a query, we searched for close relatives of UPφ901 in published bacterial genomes and assemblies. We found homologous prophages in over 200 strains of E. coli, Citrobacter koseri, Klebsiella pneumoniae, Salmonella enterica, and Yersinia enterocolitica. The most similar phages tended to come from patient urine samples or from the blood or feces of patients with urinary tract infection (UTI); more distant relatives could be found in soil, water, animal feces, and contaminated food. We refer to the collection of related filamentous phages as ‘UPφ viruses’. In several cases, identical UPφ strains were found infecting multiple bacterial genera, suggesting recent host switching events. Last, we describe the prevalence of putative UPφ viruses in metagenomes.

2. Methods

2.1 Bacterial strains

Clinical isolates of E. coli and C. koseri (see Table 1) were provided by the wolfe lab at Loyola University Chicago and were originally isolated as part of separate IRB-approved studies on the urinary microbiota of women with and without symptoms of UTIs or other urinary ailments (price et al. 2016a,b; Garretto 2019; penckofer et al. 2020). These strains and their genbank accessions (where available) are summarized in Table 1. Citrobacter koseri strains in this study have not been fully sequenced. Escherichia coli JE-1 is an IncI plasmid-bearing strain typically used to propagate phage I2-2 and was obtained from the Felix d’Herelle Reference Center for Bacterial Viruses (Université Laval, QC, Canada).

<table>
<thead>
<tr>
<th>Strain ID</th>
<th>Species</th>
<th>Has UPφ?</th>
<th>WGS accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>UMB0731</td>
<td>E. coli</td>
<td>No</td>
<td>NZ_RRWFP00000000</td>
</tr>
<tr>
<td>UMB0901</td>
<td>E. coli</td>
<td>Yes</td>
<td>NZ_PKHH00000000</td>
</tr>
<tr>
<td>UMB1220</td>
<td>E. coli</td>
<td>Yes</td>
<td>NZ_RRV00000000</td>
</tr>
<tr>
<td>UMB1526</td>
<td>E. coli</td>
<td>Yes</td>
<td>NZ_RRVH00000000</td>
</tr>
<tr>
<td>UMB8514</td>
<td>E. coli</td>
<td>Yes</td>
<td>NZ_RRUX00000000</td>
</tr>
<tr>
<td>UMB6655</td>
<td>E. coli</td>
<td>No</td>
<td>NZ_RRUL00000000</td>
</tr>
<tr>
<td>UMB6890</td>
<td>E. coli</td>
<td>No</td>
<td>NZ_RRUI00000000</td>
</tr>
<tr>
<td>UMB1389</td>
<td>C. koseri</td>
<td>No</td>
<td>Unsequenced</td>
</tr>
<tr>
<td>UMB7451</td>
<td>C. koseri</td>
<td>Yes</td>
<td>Unsequenced</td>
</tr>
<tr>
<td>UMB8248</td>
<td>C. koseri</td>
<td>No</td>
<td>Unsequenced</td>
</tr>
<tr>
<td>JE-1</td>
<td>E. coli</td>
<td>No</td>
<td>Unavailable</td>
</tr>
</tbody>
</table>

*a* UMB0901 was sequenced as ‘D75’ in its original BioProject (PRJNA316969).

*b* Citrobacter strains not sequenced at time of writing. Phage identified by PCR.

2.2 Identifying phage in bacterial genomes

UPφ901 was originally discovered as part of predicting phage sequences using phaster (Arndt et al. 2016) in UMB0901 and other E. coli isolates in a previous study (Miller-Ensminger et al. 2018; Garretto 2019). The ‘UP’ in its name originated uncreatively from ‘urine project’, though not all related viruses are necessarily found in urine, as described in Section 3.) To identify UPφ901 in other published genomes, we took advantage of its integration at its hosts’ dif sites as a tandem repeat. We defined a preliminary BLAST (Altschul et al. 1990) query sequence for UPφ901 as the region starting with the dif locus in UMB0901 and extending to the first phage repeat (7,560 bp). Throughout this paper, we will refer to the specific sequence (and homologs that are over 99 per cent identical by amino acid sequence) as ‘UPφ901’. Remaining phages that share a common set of core genes (identified below) and with less than 99 per cent amino acid identity are collectively referred to as ‘UPφ viruses’. The dif site itself is not repeated in full within the tandem duplications. We then used this single copy of the phage as a query for a blastn search using the NCBI web tool (https://blast.ncbi.nlm.nih.gov) with default parameters. This BLAST search identified prophages with over 99 per cent query coverage and over 98 per cent nucleotide identity to UPφ901 in strains of C. koseri, E. coli, K. pneumoniae, S. enterica, and Y. enterocolitica. Notably, neither PHASTER nor VirSorter (Roux et al. 2015) consistently predicted UPφ901 homologs in bacteria that had significant BLAST hits. This high false-negative rate for these two tools is likely due to UPφ901’s short genome and the reliance of these tools on information about known viruses. We have not tested the newer tool Inovirus_detector (Roux et al. 2019) with genomes carrying UPφ901 homologs.

We next downloaded all available draft and complete assemblies on NCBI for the five bacterial species listed above (as of August 2018). Each assembly was queried locally with tblastx with default parameters to identify contigs containing putative phage regions. We retained any contigs containing a hit covering at least 1,000 nucleotides of the UPφ901 query with over 75 per cent nucleotide identity. In no case did we observe a putative UPφ901 homolog on two contigs in the same assembly. The start positions of each phage were then confirmed by a separate
local blastn search for the dif site (as identified in Carnoy and Roten 2009), and stop positions were identified by confirming the first repeat in the BLAST results. These repeats typically matched just the first fifteen bases in the dif site, and we accepted putative hits that covered at least fourteen of the twenty-eight dif site nucleotides with at least 85% per cent identity. Due to varying assembly quality, not all assemblies included phage repeats, and some sequences did not include a dif repeat as an obvious stopping point. In these instances, we could only rely on the end position of the last significant hit in the initial tblastx query.

To confirm the sensitivity of our UPφ901 detection, we then blasted each of the 229 putative UPφ901-infected genomes against a query for the related phage, Ypfφ (from Y. pestis CO92, GenBank: CP000973), which has been found at dif sites in other studies (Derbie et al. 2007). We liberally accepted all putative hits, regardless of identity, and found a single case (GCA_001519645) with a cumulative query coverage over 1,000 bp. We queried both UPφ901 and Ypfφ in SRA BLAST searches against the original reads for this assembly (SRX1528813).

2.3 Phylogenetic analyses

We used Anvi’o (Eren et al. 2015) to facilitate gene clustering and annotation for the putative phage regions found in GenBank assemblies. Standard annotation tools fail to identify invovirus genes accurately, so we then used blastp through the NCBI web tool with individual phage Ike (GenBank: NC_002014) and I2-2 (GenBank: NC_001332) genes to confirm each gene’s correct annotation. These were identified by visually comparing the aligned genes rather than relying on a significance threshold, as homologous sequences may share less than 30 per cent amino acid identity in these phases. Full-length UPφ901 and most of the putative phage regions have every gene found in Ike and I2-2 (as in Fig. 2).

We identified nine UPφ901 core genes to use for phylogenetic inference (genes I, II, III, IV, V, VI, VIII, h1, and h2). Genes VII and IX were excluded from analyses, because they consist of only about thirty amino acids and were not consistently identified as genes. Gene X was also excluded as it is contained entirely within gene II. Next, we used MAFFT (Katoh and Standley 2013) to align each of the core genes. In reviewing individual gene alignments, we identified cases in genes I, II, III, and IV where the original gene calls by Anvi’o split the open reading frame (ORF) into two pieces because of early stop codons. It is unknown if these genes are truly pseudogenized or if they might still encode functional proteins. For the purpose of building the trees, we concatenated the two halves of these genes, since their complete sequences reflect evolutionary relationships, even if they might not be translated in full. Both the original and concatenated versions of these sequences are provided in the data repository for this work.

We then built a phylogeny of the 229 complete prophages with IQ-TREE version 1.6.12 (Nguyen et al. 2015) using the concatenated alignment of each core UPφ901 gene amino acid sequence with 1,000 bootstraps. We took advantage of IQ-TREE’s integrated ModelFinder (Kalyaanamoorthy et al. 2017) option to perform model selection. This step identified VT+F+R2 (variable time with empirical state frequencies and two rate categories) as the best substitution model according to Bayesian Information Criterion (BIC). Individual gene trees shown in Supplementary Fig. S1 were built using FastTree (Price, Dehal, and Arkin 2010).

We also generated a bacterial phylogeny for the phage hosts, using Anvi’o to identify a set of 400 single-copy core genes shared among each host that also contained a prophage in the virus phylogeny. These genes were aligned within Anvi’o using MUSCLE (Edgar 2004) and concatenated. We used IQ-TREE as above to build the final phylogeny. For the host tree, ModelFinder identified JTT+F+R10 (Jones–Taylor–Thornton (Jones, Taylor, and Thornton 1992) with empirical state frequencies and ten rate categories) as the best substitution model by BIC.

Trees were visualized using iTOL (Letunic and Bork 2019) and the ape package (Paradis and Schliep 2019) in R (R Core Team 2013). Additional metadata for tree visualization were obtained from NCBI, EBI, or from the literature. Key metadata included the type of material sampled for isolating bacteria (e.g. soil, blood, urine, feces) and the more general source of that material (e.g. environmental, animal, human, food). The full metadata associated with each sample is provided in the data repository (see Data availability). Nine strains had no available metadata from a publication or database. Twenty-two other strains without available metadata were from the ‘100K Pathogen Genomes Project’ (BioProject PRJNA186441: Weimer 2017) and are identified as ‘100K Project’ in figures.

2.4 Phage in SRA metagenomes

We used searchsra (Levi et al. 2018; www.searchsra.org) to check for UPφ901 relatives in metagenomes. We used a truncated 6,703 bp UPφ901 reference that included only the core UPφ viruses in phylogenetic analyses. The results from the search contained over 100,000 potential hits, most of which are false positives returned from a small number of erroneous reads aligning to the reference genome. Following the searchsra authors’ GitHub repository (https://github.com/linsalrob/SearchSRA), we first identified the depth of coverage across our reference genome for each potential hit. We used the pileup.sh function from BBMap (Bushnell 2014) to perform this step. We filtered results by both the query coverage and smoothness of coverage (Aziz et al. 2015) to identify true positives. A detailed description of these filtering steps with examples of true and false-positive coverage plots is provided in the Supplementary Methods. Metadata for the likeliest true-positive results are available in Supplementary File S1 and as part of the data repository. Statistical analysis related to these metadata was done in R.

2.5 Phage presence in culture medium

UPφ901-infected strains were grown in lysogeny broth (LB) at 37 °C with moderate shaking overnight. The following morning, 1 ml of each culture was removed, centrifuged at 16,000 g for 1 min, and the supernatants were filtered through 0.2-μm polycarbonate syringe filters. An 80-μl sample of each filtrate was then treated with ONTIZYM DNase I (Fisher BioReagents BP81071) for 30 min, followed by heat inactivation with EDTA at 65 °C for 10 min. This DNase step was included to remove any bacterial genomic DNA from the supernatant, which could result from cell death due to lysis from other prophages or shear forces. DNase-treated samples were incubated at 95 °C for 10 min to denature phage protein coats and expose the phage ssDNA. We then amplified phage genomic DNA using UPφ901-specific primers, UPphi_shortFW (GGGTATCATCAAGGGTGTCAG) and UPphi_shortRV (AGGATGCCCTAAGTCACGC). 16S PCR (63F/1387R primers) was used to confirm the absence of bacterial genomic DNA in the DNase-treated filtrates. The UPφ901 and 16S PCRs were performed on 1 μl of unfiltered culture as
positive controls. Positive UPϕ PCR products were verified by Sanger sequencing using the UPphi_shortFW primer.

### 2.6 Phage infection assays

We tested the potential for UPϕ901 to infect new hosts using standard plaque assays by spotting 5 μl of filtrate from UMB0901 (as described above) on 0.7 per cent agar LB overlays containing candidate bacterial hosts (UMB0731, UMB6655, UMB6890, JE-1).

We also tested for new infections using PCR. Colonies of prospective host strains were added to 5 ml of LB supplemented with 50 μl of filtered supernatant from UMB0901. Uninfected controls of each strain were also grown. These cultures were incubated for 18 h. Following growth, 2 μl of each culture were used in a PCR designed to test for integration into the new host’s chromosome. The forward primer (IntCheck_FW: GTGTTGATATCGAATGTCG) recognized a region early in the phage genome. This PCR can only amplify DNA if UPϕ901 integrates into a new genome. An overnight culture of UMB0901 was used as a positive control for each reaction, and the uninfected cultures were used as negative controls.

### 3. Results

#### 3.1 Initial identification

We originally identified UPϕ901 in a clinical isolate of E. coli (UMB0901) as part of separate work examining prophages present in the urinary microbiome (Garretto 2019). We then found prophages with over 99 per cent nucleotide identity in three other patient isolates in our lab collection (UMB1220, UMB1526, and UMB5814). In each case, we confirmed by PCR that the phage is shed from infected bacteria, as centrifuged culture supernatant treated with DNase is positive for UPϕ901 DNA (Fig. 1A) but negative for bacterial genomic DNA (Fig. 1B). Sanger sequencing confirmed that each PCR product was a true match to the UPϕ901 genome.

#### 3.2 Comparison to characterized inoviruses

UPϕ901 is approximately 50 per cent identical to phages IKE and I2-2 (70% coverage and ~70% amino acid identity for covered genes), each a non-integrating member of the Lineavirus genus of filamentous phages. These phages are closely related to the F-specific filamentous phage, M13. We also compared UPϕ901 to Ypfϕ, a dif-integrating filamentous phage that infects pathogenic strains of E. coli and Y. pestis. Overall, Ypfϕ shares less than 30 per cent amino acid identity with UPϕ901.

Figure 2 shows how UPϕ901’s genome compares with these other inoviruses, with each gene shaded according to its amino acid similarity with the homolog in UPϕ901. UPϕ901 contains each of the ‘core’ genes characteristic of filamentous coliphages with gene order preserved. UPϕ901 and Ypfϕ are less than 20 per cent identical in all genes except for genes II and V, which are each over 85 per cent identical between these phages. Genes II and V are both involved in regulating phage DNA replication, and this conserved module is less than 15 per cent similar in IKE and M13.

Two other differences set UPϕ901 apart from the non-integrating phages, I2-2, IKE, and M13: 1, UPϕ901 includes a copy of the E. coli dif site at the start of its genome, enabling it to integrate via the XerCD recombinase system; 2, UPϕ901 encodes two putative genes of unknown function within a region that is intergenic in I2-2, IKE, and M13. These genes are not homologous to any known protein, and all attempts to find annotated homologs of these genes in GenBank returned hits to hypothetical proteins in prophages related to UPϕ901. In addition to these hypothetical proteins, the UPϕ901 prophage region also contains three ORFs in reverse orientation just downstream of the dif site. This genome organization is mirrored by Ypfϕ and other integrating filamentous phages, including CTXϕ, where accessory genes often surround the core functions (Derbise et al. 2007; Mai-Prochnow et al. 2015).

#### 3.3 Comparative genomics

UPϕ901 strains share nearly 100 per cent amino acid identity across infected clinical isolates of E. coli in our lab collection. The only distinction is that the attachment protein, g3p (encoded by gene III), in UPϕ901 has an additional repeat of a glycine-rich motif (‘GGGES’) than the viruses found in UMB1220, UMB1526, and UMB5814. As with many other filamentous phages, UPϕ901 is integrated as a tandem repeat at the dif site in each of these genomes. In the case of UMB0901, we reassessed our own assembly by aligning the raw reads to the contig with UPϕ901 and estimating the depth of coverage (Supplementary Fig. S2). On average, the UPϕ901 prophage region has three times the coverage of the surrounding loci, indicating a tandem triple in UMB0901.

---

**Figure 1.** Phage DNA confirmation. PCR results using (A) UPϕ or (B) 16S primers after DNase treatment to remove bacterial genomic DNA from filtrates of UMB0901, UMB1220, UMB1526, and UMB5814 cultures. The control lane was from UMB0901 colony PCR.
Using the UPφ901 sequence as a BLAST query, we searched GenBank for homologous prophages in other bacteria. Prophages with over 98 per cent nucleotide identity across the UPφ901 query sequence could be found in published genomes of *E. coli*, *S. enterica*, *C. koseri*, *K. pneumoniae*, and *Y. enterocolitica*. We then downloaded all available assemblies from GenBank of these five species (26,397 in total) and used tblastx to identify viruses integrated at *dif* sites. In all, 331 genomes harbored a potential UPφ901 prophage (see Table 2 for a summary by host). We then downloaded all available assemblies from GenBank of these five species (26,397 in total) and used tblastx to identify viruses integrated at *dif* sites. In all, 331 genomes harbored a potential UPφ901 prophage (see Table 2 for a summary by host). Of these 331 putative prophages, 229 contained each gene found in UPφ901, excluding the three ORFs neighboring the *dif* site. The third of these ORFs (labeled h3 in Fig. 2) is present in 221 of these 229 putative prophages, whereas h4 and h5 are found in only 56 and 57 genomes, respectively. For the remainder of this paper, we will refer to this collection of 229 strains as ‘UPφ viruses’. The 102 remaining genomes contained a significant hit to UPφ901, but assembly or sequencing quality was inadequate to predict full phage genomes reliably. We also checked each of the 229 UPφ-infected genomes for Ypfφ to rule out possible false-positive results. We found one instance (*E. coli* assembly GCA_000457405) of coinfection between a *dif*-integrated UPφ prophage and a putative Ypfφ prophage integrated at a different locus. None of the remaining 228 genomes harbored a prophage resembling Ypfφ.

Using the UPφ901 sequence as a BLAST query, we searched GenBank for homologous prophages in other bacteria. Prophages with over 98 per cent nucleotide identity across the UPφ901 query sequence could be found in published genomes of *E. coli*, *S. enterica*, *C. koseri*, *K. pneumoniae*, and *Y. enterocolitica*. We then downloaded all available assemblies from GenBank of these five species (26,397 in total) and used tblastx to identify viruses integrated at *dif* sites. In all, 331 genomes harbored a potential UPφ901 prophage (see Table 2 for a summary by host). Of these 331 putative prophages, 229 contained each gene found in UPφ901, excluding the three ORFs neighboring the *dif* site. The third of these ORFs (labeled h3 in Fig. 2) is present in 221 of these 229 putative prophages, whereas h4 and h5 are found in only 56 and 57 genomes, respectively. For the remainder of this paper, we will refer to this collection of 229 strains as ‘UPφ viruses’. The 102 remaining genomes contained a significant hit to UPφ901, but assembly or sequencing quality was inadequate to predict full phage genomes reliably. We also checked each of the 229 UPφ-infected genomes for Ypfφ to rule out possible false-positive results. We found one instance (*E. coli* assembly GCA_000457405) of coinfection between a *dif*-integrated UPφ prophage and a putative Ypfφ prophage integrated at a different locus. None of the remaining 228 genomes harbored a prophage resembling Ypfφ.

We then built a phylogeny for the 229 complete UPφ prophages. Overall, the phylogeny reflects low genetic variation across the UPφ viruses, and most clades are separated by short branches and include polytomies. Even the more distant phage relatives found in *Y. enterocolitica* are 98 per cent identical in amino acid sequence to the original UPφ901 core genes. As a result, the tree has multiple splits with bootstrap support below 50 per cent. For this reason, we show the tree as a cladogram in Fig. 3A and will not attempt to over-interpret the tree topology. (A high-resolution phylogeny with complete bootstrap supports is shown in Supplementary Fig. S3.) For each of the genomes in the phylogeny, we identified metadata, where available, for the sample source material (e.g. urine, feces, blood) and source environment (e.g. human, animal, environmental) and added these data to the tree visualization.

UPφ901 is indicated by an arrow in Fig. 3A and is part of a clade of sixty-five phages. Fifty-five of these prophages come from *E. coli*, of which thirty-six came from urine (including fourteen UTI samples). Two large polytomies make up the majority of this group, and at least one sample (assembly GCA_000457405) was from the blood of a patient with UTI-induced bacteremia. In addition, each polytomy includes cases where identical phages (i.e. 100% amino acid identity across the genes used in the phylogeny) were also found within *C. koseri* and *K. pneumoniae*. This group contains three samples from animals. The rest of the tree includes seventy prophages found in *S. enterica* (from both animals and humans), though phages infecting each of the other host species are also present. None of these *S. enterica* samples were isolated from urine, and only ten of the remaining ninety-four sequences came from urine or UTI samples. The total counts for each host species, sample material, and source are indicated in the legend for Fig. 3.

---

**Figure 2.** Comparison of UPφ901 to related inoviruses. Roman numerals are the names for core inovirus genes common in M13 and the Lineavirus members. Ypfφ genes labeled as in Derbise et al. (2007). Genes shaded gray are darker if they are more similar to the sequence from the UPφ901 prophage in UMB0901.

**Table 2.** Prevalence of UPφ in GenBank assemblies.

<table>
<thead>
<tr>
<th>Host species</th>
<th>GenBank assemblies</th>
<th>Assemblies with UPφ hit</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>12,401</td>
<td>189</td>
</tr>
<tr>
<td><em>S. enterica</em></td>
<td>9,144</td>
<td>95</td>
</tr>
<tr>
<td><em>K. pneumoniae</em></td>
<td>4,647</td>
<td>29</td>
</tr>
<tr>
<td><em>Y. enterocolitica</em></td>
<td>177</td>
<td>10</td>
</tr>
<tr>
<td><em>C. koseri</em></td>
<td>28</td>
<td>8</td>
</tr>
</tbody>
</table>

---
We next constructed a phylogeny for the host bacteria, using a set of conserved single-copy genes identified with Anvi’o (Eren et al. 2015). The bacterial tree shows the expected breakdown by host species, with additional structure corresponding to sample material and whether the strain was from human or animal sources. Branch supports are generally higher...
in the bacterial tree except for clades with very short branches. Due to the low branch support of the phage tree, it is difficult to compare the phage and host tree topologies.

As expected from the short branches of the phage tree, individual phage gene trees (Supplementary Fig. S1) show little variation in amino acid sequence across the phages, with most genes composed of a small number of unique variants (Fig. 4). Gene III stands out as having the most variation, with ninety-five distinct amino acid sequences among the 229 genomes. Examining the alignment for gene III, most of the variation is in the number of glycine-rich repeats in the region that also varied in our lab isolates.

### 3.4 Phages in metagenomes

We next used searchsra (Levi et al. 2018) to find relatives of UPφ901 in metagenomes. The initial results included over 100,000 sequencing runs with at least one read aligned to UPφ901. After a series of filtering steps to remove erroneous data and false-positive results (see Supplementary Methods) we identified 257 SRA runs that were most likely to have true-positive results. We then used SRA BLAST (Camacho et al. 2009) to validate these putative hits and identified any instances of redundant BioSample numbers (multiple sequencing runs from the same sample) or redundant sample source (e.g. multiple samples from the same individual). This validation step identified 78 true positive, non-redundant samples that covered over 75 percent of the UPφ901 reference genome with over 90 percent sequence identity. SRA BLAST is sensitive enough to distinguish between UPφ and Ypfφ prophages (Supplementary Fig. S4).

These positive metagenomic results include sixty-four from gut, five from urine, and four from environmental samples. Two additional hits came from reference isolates of C. koseri and from mock communities used in validating microbiome projects. (Full metadata are provided in Supplementary File S1.) Five of the positive gut samples were from non-human mammals, including three from cats and two from pandas. Environmental samples included two from the New York subway, one from wastewater, and one from river sediment. Most notably, twenty-four of the sixty-four unique gut samples came from nine different studies of the infant microbiome and included children from China, Estonia, Finland, Singapore, and the USA. One hypothesis for this high number of infant samples carrying UPφ-infected bacteria is that UPφ may be more common in hospital-acquired Proteobacteria.

Several filamentous phages have been associated with pathogen virulence (Mai-Prochnow et al. 2015), including one study identifying their potential involvement in cases of neonatal meningitis caused by E. coli O18K1:H7 (Gonzalez et al. 2002). We were interested to see if any of the infant microbiome studies identified through searchsra might shed light on potential connections between UPφ viruses and bacterial virulence.

In one study (Ward et al. 2016), the authors sequenced the gut microbiomes of 144 pre-term and twenty-two full-term infants and identified cases where necrotizing enterocolitis (NEC) was associated with strains of uropathogenic E. coli. Furthermore, the anonymized patient metadata were available with the publication and could be matched to sample metadata from the SRA. This patient data included indicator variables for high (over one or ten) per cent abundance of E. coli and K. pneumoniae, as well as information on patient outcomes. We identified nine patients in this study that carried bacteria infected by a putative UPφ prophage (patient IDs: 11211, 11961, 11962, 12321, 21461, 30031, 30081, 30251, 30252). Of these, six contained E. coli and five contained K. pneumoniae; two patients had both bacteria present. The overall rate of UPφ presence (9/166) is about four times higher than across GenBank assemblies of E. coli and K. pneumoniae (218/17048). This difference represents a significant enrichment (χ² = 170.44, P < 0.001) in this study. Moreover, two of the nine patients were ultimately diagnosed with NEC. This is not significantly different (χ² = 0.0011, P = 0.973) from the rate of NEC without UPφ in the study (25/157). In addition, four of the patients (IDs starting with ‘30’) were full-term and not treated in the NICU. The rate of UPφ prophages among the full-term infants (4/22) is significantly greater than among the pre-term patients (5/144) (χ² = 5.44, P = 0.020). Thus, while it is possible that UPφ is more common in hospital-acquired bacteria, there is no evidence to suggest that it is associated with patient outcomes in this study.

### 3.5 Identifying additional infections in the lab

After identifying UPφ viruses in other hosts, we returned to the lab to test additional strains. Though only twenty-eight C. koseri genomes were available in GenBank, nearly one-third carried a UPφ virus. In two cases, these C. koseri phages were identical to prophage sequences found in urine E. coli. Given this high frequency of infection and similarity to UPφ901, we tested three unsequenced C. koseri isolates (UMB1389, UMB7451, UMB8248) using UPφ901 PCR primers. The PCR results identified an integrated phage in UMB7451, but it was not actively produced, in contrast to the infected E. coli isolates (Supplementary Fig. S5).

We also attempted to establish new infections using filtrate from UMB0901 cultures. Filamentous phages typically rely on conjugative pili as a primary receptor for infection (Mai-Prochnow et al. 2015), and each E. coli carrying UPφ901 in our lab collection also harbored an IncI conjugative plasmid. We identified three uninfected E. coli strains (UMB0731, UMB6655, and UMB8980) as candidates for testing UPφ901’s ability to establish new infections. These strains all clustered together with UPφ901-infected strains (UMB1220, UMB1526, UMB5814) in a previous gene presence-absence analysis of strains in our
collection of urinary \textit{E. coli} (Garretto 2019), contained an IncI plasmid, and were not already infected by UP\textphi901.

Despite their potential as candidate hosts, UMB0901 filtrates that were confirmed to contain the phage did not produce plaques on any of these strains. We were also unable to observe plaques on \textit{E. coli} JE-1, an IncI-bearing strain that is the standard host for UP\textphi901’s closest known relative, I2-2 (Bradley, Coetzee, and Hedges 1983). Last, we used PCR to test for phage integration in these candidate hosts but were again unable to confirm new infections.

### 4. Discussion

We have introduced a new filamentous phage, UP\textphi901, found in multiple environments, with the largest clade found predominantly in urine \textit{E. coli}. In addition to infecting \textit{E. coli}, related viruses can be found in \textit{S. enterica}, \textit{K. pneumoniae}, \textit{C. koseri}, and \textit{Y. enterocolitica}. In several cases, identical phage genomes were found in multiple host genera. These phages all share two putative genes of unknown function. We have tentatively named this collection of phages carrying these genes as ‘UP\textphi viruses’.

Three additional novel ORFs were found in UP\textphi901 and fifty-five other UP\textphi viruses, and most UP\textphi prophages lacking these genes harbored other ORFs in these positions.

These UP\textphi viruses are related to phages IKe and I2-2 in the \textit{Lineaviridae} genus of \textit{Inoviridae}, but IKe and I2-2 cannot integrate into the host genome and lack the accessory genes that appear to be unique to UP\textphi phages (Fig. 2). Given the prevalence of \textit{dif} site integration among filamentous phages and their high sequence similarity to much of the UP\textphi901 genome, it is likely that IKe and I2-2 evolved from an integrating ancestor.

Recent work has called into question the current taxonomy of filamentous phages and has suggested that the \textit{Inoviridae} may require substantial revision into multiple new families of viruses (Roux et al. 2019). We are, therefore, hesitant to claim that UP\textphi viruses deserve to be identified as a new phage genus. For the time being, we propose that the UP\textphi viruses should be considered members of the existing \textit{Lineaviridae} genus, with the current members (IKe and I2-2) forming a subclass of phages that have lost the \textit{dif} site and nearby accessory genes. We acknowledge, though, that this taxonomy could change as the \textit{Inoviridae} are revised and may be complicated by recombination between co-infecting phages.

#### 4.1 Relationship to CUS\textphi and Ypf\textphi

Prior work identified the filamentous prophages CUS\textphi and Ypf\textphi in strains of extra-intestinal \textit{E. coli} (ExPEC) and \textit{Y. pestis} (Gonzalez et al. 2002; Derbise et al. 2007). While these phages have variable accessory gene regions (the same regions where UP\textphi has its own novel genes), CUS\textphi and Ypf\textphi otherwise share 99 per cent nucleotide identity with each other (Derbise et al. 2007). These phages are found in similar hosts as UP\textphi viruses and rely on \textit{dif} site integration. In these prior studies, the authors demonstrated both the potential for these phages to increase the virulence of ExPEC (Gonzalez et al. 2002), and also the likely association between Ypf\textphi and the modern strains of \textit{Y. pestis} pv orientalis associated with the third plague pandemic (Derbise et al. 2007).

As part of the exploration of UP\textphi, we compared the individual genes from our UP\textphi901 reference sequence to the Ypf\textphi prophage region in its type host, \textit{Y. pestis} CO92. We found that only genes \textit{II} and \textit{V} have significant homology (Fig. 2), whereas the rest of UP\textphi901 is more closely related to phages I2-2 and IKe. Genes \textit{II} and \textit{V} are both involved in regulating phage DNA replication, and this modular split in homology between the disparate viruses likely reflects a deeper history of recombination among the phages.

Because they infect related hosts and might have a history of recombination, we also checked for any instances of co-infection between UP\textphi and Ypf\textphi in our data. We found one such case (GCA_001519645), corresponding to \textit{E. coli} bacteremia in a patient with sepsis. This assembly, unfortunately, consisted of 206 contigs, and Ypf\textphi appears to be split across at least four different contigs. Notably, Ypf\textphi is not included on a large contig containing both \textit{dif} and UP\textphi, suggesting this Ypf\textphi prophage integrated at a different locus using another mechanism or is present as a plasmid in this host. SRA BLAST (Supplementary Fig. S6) confirmed that each phage is fully covered by the reads. It is unknown if the prophage genomes interact. Sachs and Bull (2005) demonstrated that co-infecting filamentous phages (in their case, f1 and IKe) could coevolve to package their genomes together using shared coat proteins. It is feasible that UP\textphi and Ypf\textphi could also produce chimaeric virions. In fact, the UP\textphi region in GCA_001519645 has an early stop codon in its gene III and would likely be unviable unless it is able to incorporate Ypf\textphi’s version of the attachment protein during assembly.

#### 4.2 Tandem repeats and prophage diversity

\textit{dif} site integration typically results in tandem duplication of the phage genome (Mai-Prochnow et al. 2015). It has been observed in CTX\textphi and Ypf\textphi that hosts harboring a \textit{dif}-integrated phage can be super-infected by other such viruses, with the new phage potentially supplanting the original (Midonnet et al. 2019) or integrating in the same region (Derbise and Carniel 2014). In the latter cases, the phage can integrate between the original copies or downstream (Chouikha et al. 2010). It is, therefore, possible to observe tandem triples or even quadruples of inovirus prophages at the \textit{dif} site of infected bacteria, and these downstream homologs are not always identical. Furthermore, this series of integration events can promote hybridization between co-infecting inoviruses when new phages are produced by infected hosts (Davis and Waldor 2000; Derbise and Carniel 2014), as well as co-packaging of phages into shared coats as described above (Sachs and Bull 2005). In the case of UP\textphi prophages, we did not always know the exact number of tandem repeats within each host genome, as most sequencing efforts relied on only short Illumina reads that were not resolved into tandem copies during assembly by the original data providers. For UMB0901, we used depth of coverage to infer that UP\textphi901 is integrated as a tandem triple.

Given the propensity for tandem prophage recombination, there is additional uncertainty in interpreting and comparing the genomes of UP\textphi viruses that presents an additional challenge for taxonomy. The \textit{E. coli} strain ECONIH2 (Gen\textit{Bank}: CP014667) was originally sequenced with both Illumina and PacBio technologies and offers a valuable case study. ECONIH2 contains a tandem triple of UP\textphi prophages. The first two copies of the phage are identical by nucleotide sequence and contain each of the genes also found in UP\textphi901, as well as three additional ORFs. The third copy of the phage in ECONIH2, however, contains only the first of these additional genes followed by a fourth novel ORF. This one bacterial strain demonstrates the potential heterogeneity among co-infecting prophages.
4.3 Future work
Additional work remains to understand the prevalence and role of UPφ viruses in different microbial communities. First, future research will need to characterize the two novel genes, h1 and h2. These genes are found in a region of the genome that often includes genes that alter host behavior or virulence (Mai-Prochnow et al. 2015). In the prototypical integrating inovirus, CTXφ, this region encodes the cholera toxin genes and their regulators (Waldor and Mekalanos 1996). It appears likely that these two new genes interact with one another, but they share no homology to any known gene, and there is little to hint at their possible functions.

While we have identified putative phages from genome assemblies, it was not always possible to predict phage end positions or the number of tandem repeats. These issues make it difficult to assess the pan-genome of UPφ viruses, as some might carry additional accessory genes not found in UPφ901. Future research that includes long-read resequencing of infected hosts and phages released into culture medium would help to resolve these ambiguities.

We also attempted to determine the host requirements for establishing new infections of UPφ901. We were able to confirm by PCR that new virions are released into the culture medium by infected bacteria, but we were unsuccessful in our attempts to infect new hosts. Our failure to observe new infections may be the result of lacking appropriate host strains for testing or due to low efficiency of establishing new infections. At the same time, UPφ901’s closest relatives, I2-2 and IKe, each infect E. coli harboring IncI conjugative plasmids, and each E. coli strain infected by UPφ901 in our collection also contained an IncI plasmid. It, therefore, appears likely that at least some of these phages rely on IncI-encoded pili, and future work with additional strains and different growth conditions will test this possibility.

It is also possible that different UPφ viruses rely on different conjugative pili to initiate new infections. The phage attachment protein, g3p, showed the greatest variation across strains, and each of these variants might correspond to different host specificity. Most of this variation was observed in a glycine-rich repeat region that links N- and C-terminal domains of g3p. Previous work showed that this region can diversify in phage IKe in a few generations in the lab, but no effects on phage fitness were observed (Bruno and Bradbury 1997). Those experiments, however, did not test for changes in host range.

5. Conclusion
UPφ viruses represent an exciting new group of filamentous phages. Many inoviruses play important roles in modifying bacterial pathogens within eukaryotic hosts. It remains to be seen if UPφ viruses, though frequently found as prophages in bacteria from urine, affect the frequency of UTI or other aspects of urinary health. In the case of pre-term infant microbiomes, we observed no correlation between UPφ prophage presence and patient outcomes. It is possible that UPφ phages affect bacterial ecology within the urinary environment without changing bacterial virulence. Perhaps more important will be understanding how frequently the phages shift hosts within different environments and whether these new infections alter the phenotypes of multiple genera within a community.

Data availability
Data from this work is available at figshare (https://figshare.com/s/23504b35ae0617dac68).

Supplementary data
Supplementary data are available at Virus Evolution online.

Acknowledgments
We are grateful to the Wolfe Lab for generously providing strains of E. coli and C. koseri used in this work and to Putonti Lab members for helpful comments on the manuscript. We also thank O. Pybus, M. Marston, and two anonymous reviewers for constructive feedback and suggesting a deeper analysis of the metagenomic data and comparison to Ypfφ.

Funding
This work is supported by NSF (1661357 to C.P.).

Conflict of interest: None declared.

References


