

Reviewed Preprint

v1 • January 12, 2026

Not revised

Reviewed Preprint

v2 • June 24, 2026

Revised by authors

✉ For correspondence:

camilla.do@bcm.edu

Competing interests:

Baylor College of Medicine has filed for intellectual property protection on behalf of authors CD, PN, and AM on material related to the device in this manuscript. Remaining authors declare no competing interests.

Funding:

See [page 19](#)

Reviewing editor:

María Mercedes Zambrano, CorpoGen, Colombia

© 2026, Do et al. This article is

distributed under the terms of the

[Creative Commons Attribution](#)

[License](#), which permits unrestricted

use and redistribution provided that

the original author and source are

credited.

Pathogen-Phage Geomapping to Overcome Resistance

Camilla Do¹✉, Keiko C Salazar^{1,2}, James D Chang¹, Justin R Clark^{1,2}, Austen L Terwilliger^{1,2}, Paul Ruchhoeft⁴, Paul Nicholls³, Anthony W Maresso^{1,2}

¹Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, United States • ²TAILOR

LABS, Baylor College of Medicine, Houston, United States • ³Section of Infection Diseases, Department of Medicine,

Baylor College of Medicine, Houston, United States • ⁴Department of Electrical & Computer Engineering, University of

Houston, Houston, United States

eLife Assessment

This **important** study establishes an environmental sampling workflow for the discovery of bacteriophages capable of infecting antibiotic-resistant pathogens. The authors **convincingly** demonstrate the effectiveness of the approach, even with the limited sampling scheme and the current challenges in viral taxonomy. This study will interest researchers working on bacterial infections, environmental microbiology, and phage-based alternatives for addressing antimicrobial resistance.

<https://doi.org/10.7554/eLife.109259.2.sa3>

Abstract

The rise of antibiotic resistance has renewed interest in bacteriophages as therapeutic alternatives. However, co-evolution of phage and bacteria will naturally give rise to phage-resistant pathogens, complicating phage therapy efforts. A critical bottleneck in the production of phage therapeutics is the discovery of virulent phages against resistant pathogens. Conventional methods for discovery are time-consuming, biased, and laborious, limiting the potential for identifying suitable phage candidates.

To overcome these limitations, we combined small-volume environmental sampling with 16S rRNA sequencing to identify reservoirs where bacterial hosts co-exist with their phage predators. This strategy, which we term geographical phage mapping (geΦmapping), pinpoints ecological “hotspots” for targeted phage hunting. We further developed a portable phage hunting device (ΦHD) that generates highly enriched phage concentrates directly from these reservoirs. By integrating geΦmapping with high-throughput enrichment, we constructed the RΦ library, a diverse collection of novel phages. We captured and isolated 36 new phages targeting extremely-resistant organisms across various ESKAPE pathogens when conventional phage hunting and experimental evolution approaches failed.

Introduction

The rise of antibiotic resistance has brought global healthcare to the precipice of a post-antibiotic era, driving interest in bacteriophage (phage) therapy as an alternative strategy to combat resistant bacterial infections.^{1–3} Unlike antibiotics, phages evolve with their host; both locked in a relentless evolutionary arms race. Bacteria and phage continuously expand their repertoires of resistance and counter-resistance strategies. Among many examples, phages counter bacterial defenses by exposing receptors with enzymes, evading restriction enzymes with DNA base modification and evolving anti-CRISPR proteins.^{4,5} In response, bacteria resist phage infection by blocking adsorption and nucleic acid injection and using restriction enzymes and CRISPR-Cas9 to target phage nucleic acids.^{6,7} This dynamic interplay of phage resistance poses a barrier to

effective phage therapy.⁸ TAILOR is a phage center dedicated to providing lytic phages to physicians to treat patients battling multidrug-resistant infections.⁹ Across 263 distinct cases encompassing 462 bacterial isolates, 22% of isolates given to TAILOR remain untreated despite an extensive library of over 325 phages targeting 15 bacterial species (Fig. 1A). Information regarding TAILOR's phage and bacterial library can be found in Supp. Tables 1-3. When no phages in the library can target a pathogen, TAILOR turns to directed evolution (to expand host range of a given phage) or environmental sampling (to find natural phages). Such approaches often fail to yield a suitable phage and so discovering new, virulent phages that can effectively target these pathogens becomes a critical and rate-limiting step (Fig. 1B). As phage therapy spreads, resistant isolates will inevitably emerge under increased selection pressure, analogous to antibiotic resistance after introduction of penicillin.¹⁰ This puts the onus onto phage discovery where improvements are critically needed.¹¹

Phage engineering is one method of constructing phages for improved therapeutic potential. Examples include generating lytic derivatives of lysogenic phages and altering phage tail fibers to adjust their host range.^{12,13} In contrast, phage-host competition has naturally addressed phage resistance over billions of years of evolution. The challenge lies in isolating these phages from nature. Conventional methods for phage discovery can be time-consuming or introduce bias.¹¹ Phage hunters are limited to sampling small volumes of water, a constraint that typically necessitates the use of enrichment techniques to increase the likelihood of phage detection.¹⁴⁻¹⁷ Enrichments can favor fast-growing or highly virulent phages, potentially overlooking those that may be therapeutically relevant but less abundant or slower to propagate.¹⁸ To sample larger volumes, dead-end filtration (limited by low filtration speed) and iron chloride flocculation (limited by phage inactivation) are typically used (Supp. Table 4).¹⁹⁻²⁹ The speed and phage bias limitations are acceptable for metagenomic approaches but unhelpful for phage therapy, leaving a critical gap for patients that urgently need phages for phage-resistant bacterial infections.

Viral metagenomics involves identification of largely uncultivated viral genomes from the environment. Unfortunately, analysis is hindered by the vast diversity of viruses, lack of consensus sequences, and limited reference genomes available.³⁰ Some 70% of all viral genomes are unclassified, often referred to as “viral dark matter.”³¹ Studying unknown viral genomes can advance viral ecology and may contribute useful gene products for research and therapeutics.³² Viral metagenomics are frequently limited by low biomass, sampling biases, and limitations in sampling remote sites.^{33,34}

In this study, we sought to improve current methods for phage discovery to tackle phage-resistant pathogens and enhance gene discovery. We used a bioprospecting approach, collecting samples in East Texas, USA, and applied 16S rRNA analysis to identify pathogenic reservoirs and hence their phages. We dubbed this process, “Geographical Phage (Φ) Mapping,” or “GeΦmapping.” After identifying pathogen-rich sources, we deployed a high-throughput phage hunting device, ΦHD, for sample processing (Fig. 1C,D). We used this approach of combining molecular bioprospecting and high-volume sampling to address problematic “extremely phage-resistant organisms” (XΦROs) in clinical cases.

Results

GEΦMAPPING of water sources from Houston and Galveston, Texas

Phages are found where their hosts thrive.³⁵ In marine ecosystems, for example, phage-to-bacterium ratios often reach 10:1, reflecting close associations between bacterial density and phage populations.³⁶ We identified sites of high pathogen abundance to determine where to hunt for phages. We sampled 15 waterways around Houston and Galveston, Texas, USA thrice and performed 16S rRNA analysis to examine the bacterial constituents (Fig. 2A). We performed rarefaction analysis to estimate species richness relative to sampling effort. The logarithmic shape of the rarefaction curve indicates that most of the abundant taxa present have been detected in each sample, although sample saturation was not fully achieved (Supp. Fig. 1A, top). To further

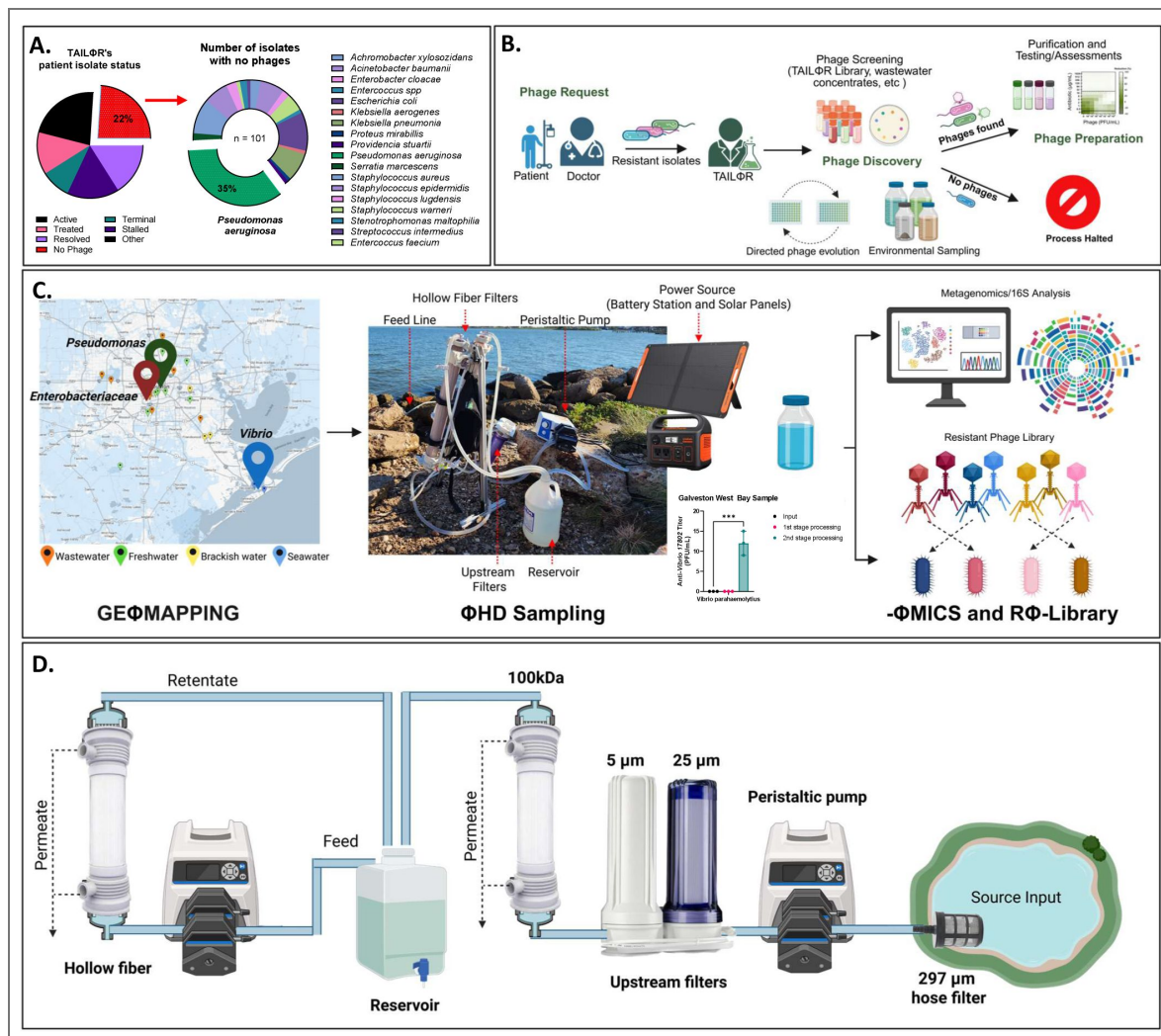


Figure 1. Overview and schematic of ΦHD.

A) Pie charts of patient isolate status and percentage of bacterial strain with no phages for in TAILΦR's library. Unfortunately, 24% of isolates that TAILΦR receives have no phages; at 35%, most of the isolates are *Pseudomonas aeruginosa*. **B)** Diagram from phage request to phage discovery and limitation. Clinicians and their patients can seek phage therapy for antibiotic-resistant infections when there are no other approved treatments available. After approval, clinicians send patient isolates to TAILΦR. TAILΦR's phage library and wastewater concentrates are screened against the patient isolate. If phages are discovered, they are moved onto the next stage for preparation where phages are ultimately purified and tested for safety and efficacy before being handed to the clinician. When there are no phages for that isolate in the library, TAILΦR searches for phages through environmental sampling. Alternatively, they can train related phages to infect the isolate through directed phage evolution. **C)** Diagram of strategies employed in our study. Geographical phage (Φ) mapping (geΦmapping) allowed us to pinpoint target sites rich in our pathogen of interest from our environment (left). Once a location is chosen, we used our capture device, the phage hunting device (ΦHD), to filter, concentrate, and enrich our sample with phage and bacteria (middle). With the samples, we further analyzed them with phage metagenomics (ΦMICS) and screened, isolated, and characterized phage present in the samples(right). This led to the curation of the resistant phage (Φ) library (RΦ-Library). **D)** Schematic of ΦHD. The components on the right of the reservoir draws, filters, and concentrates water through ΦHD. On the left of the reservoir, the remaining component concentrates any material in the reservoir. Created with BioRender.com.

quantify species richness, we compared α -diversity indices (inverse Simpson, Shannon, and Chao) and noticed that White Oak Bayou was the richest (Supp. Fig. 1B). To compare the membership of each site's microbiome, we analyzed β -diversity using principal coordinate analysis (PCoA) based on Bray-Curtis dissimilarity (Fig. 2B,C). Samples of seawater, sewage, and freshwater clustered distinctively while samples from brackish water mingled between freshwater and seawater sites (Fig. 2B,C). We performed an analysis of molecular variance (AMOVA) to compare samples derived from brackish, sea, fresh, and sewage. AMOVA revealed significant structuring across all habitat comparisons, with among-group variance exceeding within-group variance in every case. Differentiation was strongest between fresh and sewage samples ($F_s = 28.71$, $P < 0.001$), followed by sea–sewage ($F_s = 20.21$, $P < 0.001$) and brackish–sewage ($F_s = 15.92$, $P < 0.001$), indicating pronounced divergence associated with wastewater. Comparisons involving brackish with seawater ($F_s = 4.65$ – 13.16 , $P \leq 0.006$) and brackish–fresh ($F_s = 5.96$ – 13.16 , $P \leq 0.001$) were also significant ($F_s = 4.65$ – 13.16 , $P \leq 0.006$). Interestingly, a freshwater sample from Vince Bayou, where *Acinetobacter* was prominent, clustered with sewage samples suggesting its microbial profile resembled that of sewage. Of note, a contaminated industrial facility, formerly an oil processor and wastewater treatment plant, is situated upstream of this location.³⁷ After heavy storms, toxic wastes have reportedly contaminated the area with reports of hundreds of dead fish.³⁸ During sampling, toxic chemical contamination may have shifted the microbiome to resemble sewage-like communities, causing resilient genera like *Acinetobacter* to dominate.

To focus on pathogens, we extracted the counts of pathogen-related taxa and generated heatmaps normalized by site (Fig. 2D) and by taxa (Fig. 2E). By highlighting selected pathogenic taxa, we can compare patterns across the environments relative to the other taxa; this does not imply absolute abundance or general enrichment of these taxa in the environment. Genus-level classification includes both pathogen and non-pathogenic species. Normalizing by site, we determined the microbial composition of each location. We detected *Vibrio* in Gulf of Mexico and West Bay; both expected in marine environments (Fig. 2D). We noted *Enterobacteriaceae* was most common in waterways draining rural and suburban areas, while *Pseudomonas* was more common in larger urban waterways (Brays and Buffalo Bayous) (Fig. 2D, black rectangles). Again, Vince Bayou was an outlier, highly enriched with *Acinetobacter*.

Normalizing by taxa allowed us to find sites of interest for Φ HD sampling (Fig. 2E, black rectangles). The data showed predominance of *Vibrio* at marine sites, while *Pseudomonas*, *Enterobacteriaceae* and *Enterococcus* were best represented at Brays Bayou. Vince Bayou remained enriched with *Acinetobacter*. Of the 22% of TAIL Φ R isolates without a phage, 35% of these were *Pseudomonas* (Fig. 1A, Supp. Table 3). This made Brays Bayou, rich in *Pseudomonas*, an ideal target for high volume sampling.

Environmental sampling freshwater and seawater with Φ HD

Although phages are abundant in nature, harvesting phages against pathogens can be difficult.^{14,17} Typically, phage hunters are restricted to small volumes and sometimes rely on enrichment techniques to compensate.^{14,17} These techniques vary by sample and host but take days with no guaranteed success. After selecting our sampling location, the next challenge was obtaining a highly concentrated sample through high-volume sampling to ensure sufficient phage abundance for effective downstream screening.

We approached this using Φ HD, which was inspired by the various research groups that used tangential flow filtration (TFF) for viral metagenomics.^{25,26,39} We designed Φ HD to have a robust filtration train combined with a high-flux concentration component, allowing for low shear fluid path and high portability (Fig. 1D). We used a 297 μ m hose filter to screen out larger debris, followed by sequential 25 and 5 μ m filter cartridges to remove larger microorganisms such as plankton and protists. We then used two hollow fiber TFF units, operating in parallel for concentration with flow rates in the low shear region. The first unit performs ultrafiltration and fills the reservoir with retentate, while the second takes from that reservoir and continuously concentrates the retentate in a closed loop. These are powered by a dual-headed peristaltic pump equipped with a lithium-ion battery and solar panels to allow prolonged operation in the field.

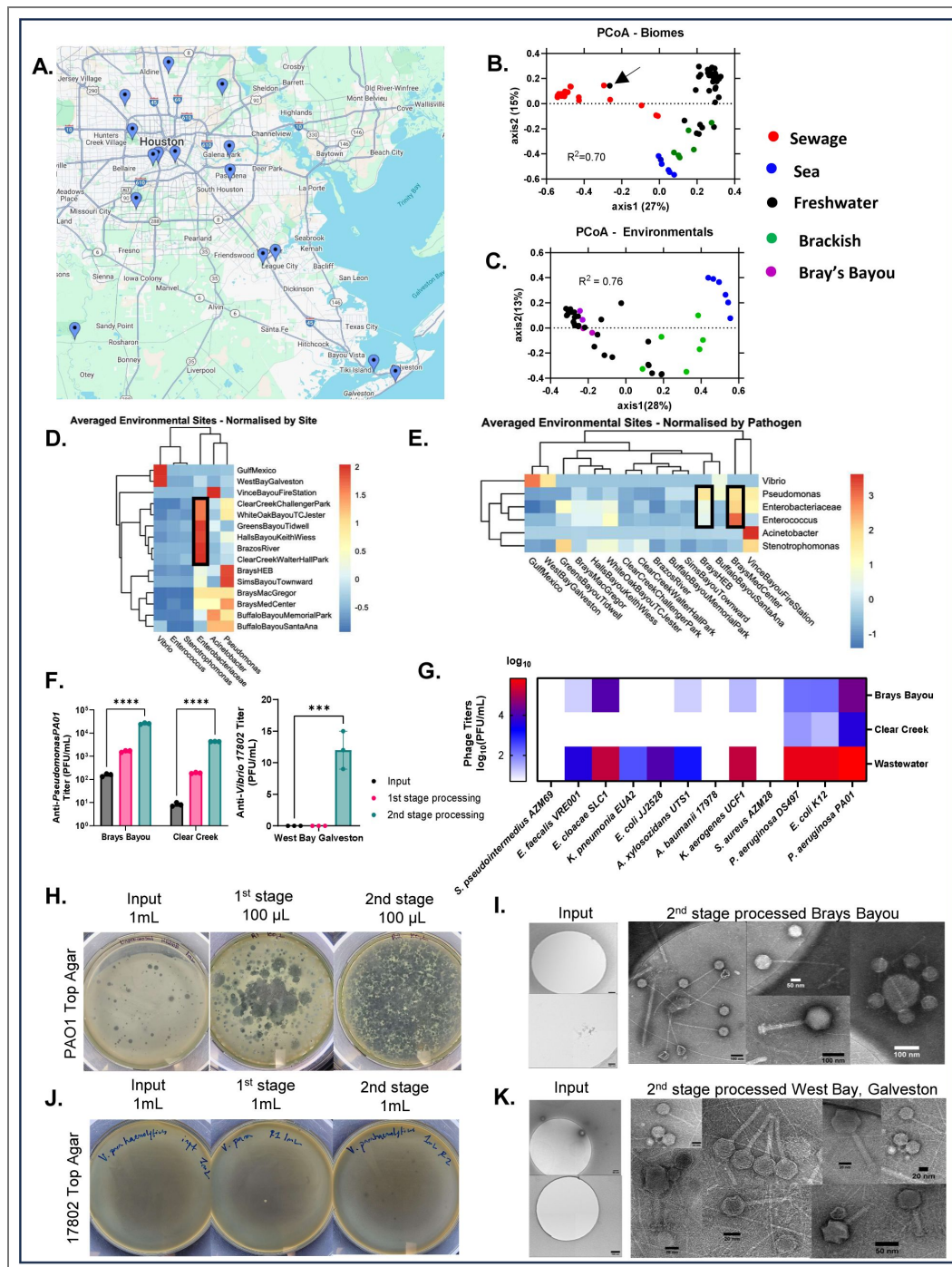


Figure 2. GEOMAPPING highlights Bray's Bayou as a *Pseudomonas* rich target site, and performance of Φ HD sampling of several environmental sites.

A) Geographical map of fresh water samples obtained in Houston, TX, USA. **B)** Principal Coordinate Analysis (PCoA) analysis showing β -diversity differences between sewage, sea, and freshwater sources. The black arrow indicates the Vince Bayou sample. **C)** PCoA analysis (Bray-Curtis dissimilarity) of environmental water bodies with Bray's Bayou target sites highlighted (purple). **D)** Heat map with z-score normalized by site shows distinct patterns of microbial prevalence at various sites. **E)** Heat map, z-score normalized by pathogen, highlights the potential of Bray's Bayou as a target site for *Pseudomonas* and enteric phages. **F)** Quantification of anti-*Pseudomonas* (left) and anti-*Vibrio* (right) plaques (PFU/mL) from different stages of processing (input, first stage, and second stage processing). **G)** Heat-map showing anti-pathogen phages ($\log_{10}$ (PFU/mL)) from freshwater sites. We display the mean and SEM of 3 technical repeat measurements. **H)** Plaque assay on *P. aeruginosa* PAO1 of unprocessed, first stage concentrate, and second stage concentrate of a freshwater sample. **I)** TEM of unprocessed water from freshwater. **J)** Plaque assay on *V. parahaemolyticus* 17802 of unprocessed, first stage concentrate, and second stage concentrate seawater sample. **K)** TEM of unprocessed water from seawater, respectively. Representative images shown (H-K).

During spike-in studies, we spiked pond water (60L) with phage (1×10^3 PFU/mL of Φ JB10) (Supp. Fig. 2A). We found minimal losses of phage and bacteria during passage through the filtration train and obtained an average 20-fold increase in phage yield (Supp. Fig. 2B). After the spike-in study, we evaluated our cleaning-in-place protocol (Supp. Fig. 2C, Post CIP) for the TFF units and found it effectively removed phages and bacteria, confirming Φ HD is reusable.

In the field, we routinely achieved flow rates of 200L/hr and gathered samples ranging from 400–1200L (Supp. Table 5). Initial testing on-site proved the filtration/concentration train was robust enough to manage the highly polluted water ways of Houston (Supp. Fig. 2D–E, Santa Ana Capture Site (SACS)). Although not all *Pseudomonas* phages will be detected, we selected *Pseudomonas aeruginosa* PA01 as the indicator strain because of its broad susceptibility profile and its role as a permissive host. After first stage processing with Φ HD, phage yield increased by 95-fold in comparison to unprocessed sample (Supp. Fig. 2E, right panel). We also sampled the more pristine waters of Pedernales Falls (PF) and Hamilton Pool Preserve (HP), and incorporated a secondary in-lab concentration stage to further increase retentate concentrations (Supp. Fig. 2F). Phage yields from inputs from PF and HP were near zero but increased to 10^{2-3} PFU/mL with secondary processing. Similar treatment of SACS retentates yielded an average 3,501-fold increase in phage yield (Supp. Fig. 2G). This shows Φ HD, with second stage concentration, greatly increased the detection of viable phages from a given site (Supp. Fig. 2G, 2H). Screening with other bacteria showed more phages against other pathogens from polluted SACS water compared to pristine sites (Supp. Fig. 2I).

To test the combination of ge Φ mapping and Φ HD, we selected one site that was highlighted by our map as *Pseudomonas*-, *Enterobacteriaceae*- and *Enterococcus*-rich (Brays Bayou, Fig. 2E) and one low-richness site (Clear Creek). We processed 400L/site and tracked phage yields by endogenous anti-*Pseudomonas* (PA01) phages. Φ HD significantly increased phage yields at each site (166-fold at Brays Bayou and 520-fold at Clear Creek, Fig. 2F,H). Using second stage concentrates, we screened a panel of patient isolates of different genera to compare hits and found more lytic phages at Brays Bayou (7/12 species) than Clear Creek (3/12 species), phenotypically validating our 16S-based mapping strategy (Fig. 2G). We employed an identical filtration train at Galveston Bay sampling the marine, saltwater biome. We tracked concentrations with *Vibrio parahaemolyticus* 17802 and could only identify plaques after second stage concentration (Fig 2F-right panel, J), showing utility in salt water. To test whether phage concentration increased without host-selection bias, we used transmission electron microscopy (TEM). We observed a wide range of phage morphologies with intact tails in concentrates, while the input was mostly clear of particles (Fig 2I,K). This provides further objective evidence of unbiased concentration of intact phages through the Φ HD process.

GE Φ MAPPING of wastewater treatment plants in Houston, Texas

Pathogens belonging to *Enterococcus* and *Enterobacteriaceae*, including *Escherichia* and *Klebsiella*, are known inhabitants of wastewater, so we reasoned wastewater would be prime choice for finding phages against these pathogens.^{40,41} In a similar fashion to freshwater mapping, we sampled influents of seven different wastewater treatment plants (WWTPs) twice to generate ge Φ maps of the pathogenic landscape of wastewater plants. Fig. 3A shows the catchment area for each WWTP in Houston. The logarithmic shape of rarefaction curves showed we adequately sampled from each site (Supp. Fig. 1A, bottom). PCoA of β -diversity (Bray-Curtis dissimilarity) demonstrated close clustering of sewage samples despite substantial diversity in both human populations and localization of industrial and commercial facilities. Unexpectedly, influent samples collected from Alameda Sims WWTP showed unusual abundance of *Chloroflexi*, known to be important in biodegradation in wastewater bioreactors (Fig. 3B, yellow).⁴²

We normalized by taxa and generated heatmaps to make our wastewater ge Φ map (Fig. 3C). The Park Ten WWTP influent was rich in *Pseudomonas*, *Acinetobacter* and *Stenotrophomonas*, but it was being decommissioned during our study. West University WWTP had higher levels of *Enterobacteriaceae* and *Enterococcus*, so we selected it for further investigation. The wastewater treatment process alters the microbiome of wastewater, so we compared the pathogenic taxa

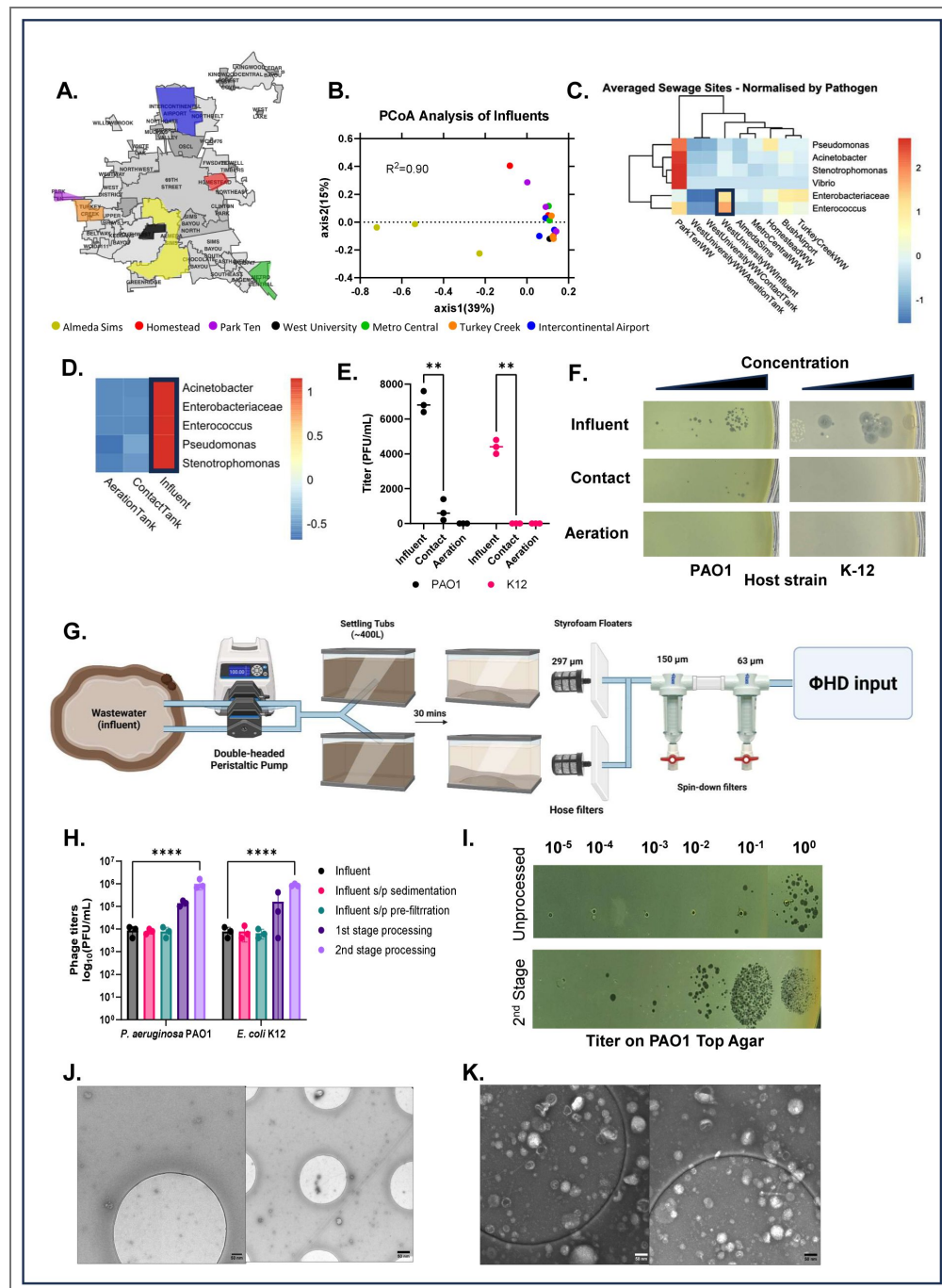


Figure 3. GEΦMAPPING identified *Enterococcus* and *Enterobacteriaceae* was rich at West University WWTP, and performance of ΦHD Sampling in wastewater.

A) Schematic of geographical areas drained by WWTPs in Houston, TX, USA. **B)** β -diversity analysis shows variation in OTUs present in WWTPs by PCoA analysis based on Bray-Curtis dissimilarity (**AB**, colors represent differing WWTPs). **C)** Heatmap analysis of pathogenic taxa highlights (black box) location of enteric pathogens (z-score normalized by taxa). **D)** 16S analysis further localizes pathogenic taxa to wastewater influent (z-score normalized by taxa). **E)** Quantitation of plaques on index strains ($n=3$, $** = p\text{-value} \leq 0.005$ by 2-way ANOVA with Tukey's multiple comparisons correction) confirms 16S sequencing data. **F)** Titration of VLPs from different WWTP processing areas on PAO1 (*P. aeruginosa*) and K-12 (*E. coli*) index strains (representative data shown). **G)** Outline of prefiltration and sedimentation step for wastewater sites. **H)** Quantification of anti-*Pseudomonas* plaques and anti-*Escherichia* plaques from different stages of processing show phage retention through the system and concentration of the end products. **I)** Titration of VLPs from different WWTP processing areas on PAO1 (*P. aeruginosa*). **J)** TEM of unprocessed ΦHD input from wastewater. **K)** TEM of second stage concentrated wastewater. Representative images shown (F, I-K).

throughout the process. We sampled raw wastewater filtered only through large steel bars that remove large debris (influent). Similarly, we sampled wastewater mixed with activated sludge (contact tank) and wastewater exposed to activated sludge with air bubbling (aeration tank).⁴³ α -diversity indices showed that microbial diversity was richest in the contact tank and aeration tank of West University WWTP (Supp. Fig. 1B).⁴³ but the influent tank had the highest abundance of the selected pathogenic taxa (Fig. 3D). We confirmed this 16S genomic data by comparing titers of phages against indicator strains (PAO1 for *Pseudomonas* and K-12 for *E. coli*) which showed the majority of these phages were found in the influent (Fig. 3E-F).

Wastewater sampling with Φ HD

We used a modified Φ HD method to process 400L of influent wastewater from West University WWTP (Fig. 3G). To avoid pump issues caused by gravity, we lifted the wastewater 1–2 meters onto an elevated deck into large tubs. We allowed this material to settle for 30 minutes to sediment large debris and then we pre-filtered the sample using 297 μ m hose filters and two spin-down sediment traps (150 and 63 μ m screens) to produce a clarified product. This pre-filtered material was then used as the Φ HD input. To ensure these additional filtration steps did not capture significant quantities of phage, we titered the material at each stage against indicator strains (PAO1 and K12) and found negligible losses (Fig. 3H). We titered input and second stage concentrates (Fig. 3H,I) on *Pseudomonas* and *Escherichia* indicators with final phage yields increasing by 118-fold and 108-fold, respectively. We attempted to look for non-selective concentration of virus-like particles (VLPs) with TEM, however this was complicated by residual particulate matter. Nonetheless, we were able to appreciate some phages in retentates but not the unconcentrated input material (Fig. 3J,K). We also screened the second stage concentrate against the previous panel of clinical and laboratory strains. This yielded phages for 9 of 12 bacterial strains screened, and phage titers were higher than at other sites (darker red) (Fig. 2G).

Collective shallow shotgun metagenomic sequencing

Phage diversity varies significantly across different biomes due to factors such as host availability, ecological niches, and environmental conditions. To test the utility of Φ HD for metagenomic applications, we compared viral metagenomes extracted from an unprocessed sample (5L) and a Φ HD concentrate (60mL of 6667X retentate, corresponding to 400L of processed sample). The majority of α -diversity indices were similar, suggesting comparable within-sample diversity; however, β -diversity based on Bray–Curtis dissimilarity (0.6062) indicates highly moderate differences in viral community composition. (Supp. Table 6). Heatmap of viral population abundances showed no clear clustering or distinct patterns between the two samples (Supp. Fig. 3A). Nevertheless, we detected substantial increases in unique viral contigs (Supp. Table 7, Supp. Fig. 3B,C). We compared putative *Pseudomonas*-infecting viruses from the input and Φ HD samples, and the Φ HD sample contained both different members from genetically distinct groups and a greater absolute number of unique anti-*Pseudomonas* viruses (Supp. Fig. 3D,E,F).

Given the apparent utility of Φ HD in concentrating samples for metagenomics, we looked at the viromes of fresh (Brays), brackish (Clear Creek), waste (West University WWTP), and seawater (Galveston). Based on α -diversity indices, either freshwater or wastewater samples represent the richest sources of viral diversity (Supp. Table 8). Heatmap of viral abundance and β -diversity analysis with PCoA confirm distinct viral community compositions between biomes (Supp. Fig. 4). We clustered our putative, non-redundant viral sequences into vOTUs, or virus operational taxonomic units. We generated a gene-sharing network with vCONTACT2 with each node representing a vOTU, colored by source and clustered them with reference genomes from the Prokaryotic Viral RefSeq v201 database (Fig. 4A,B). We counted 16,198 vOTUs $\geq$ 5kb and 5095 vOTUs $\geq$ 10kb.

39.1% of vOTUs clustered with a reference genome at the genus level, while 20.3% were outliers sharing only 1–2 genes with other genomes. 30.3% were singletons with no genes related to any other genomes in the data set (Fig. 4A,B). We also used PhaGCN to classify each virus, but only 33.5% of viruses could be classified to the family-level (Fig. 4C). Of the classified genomes,

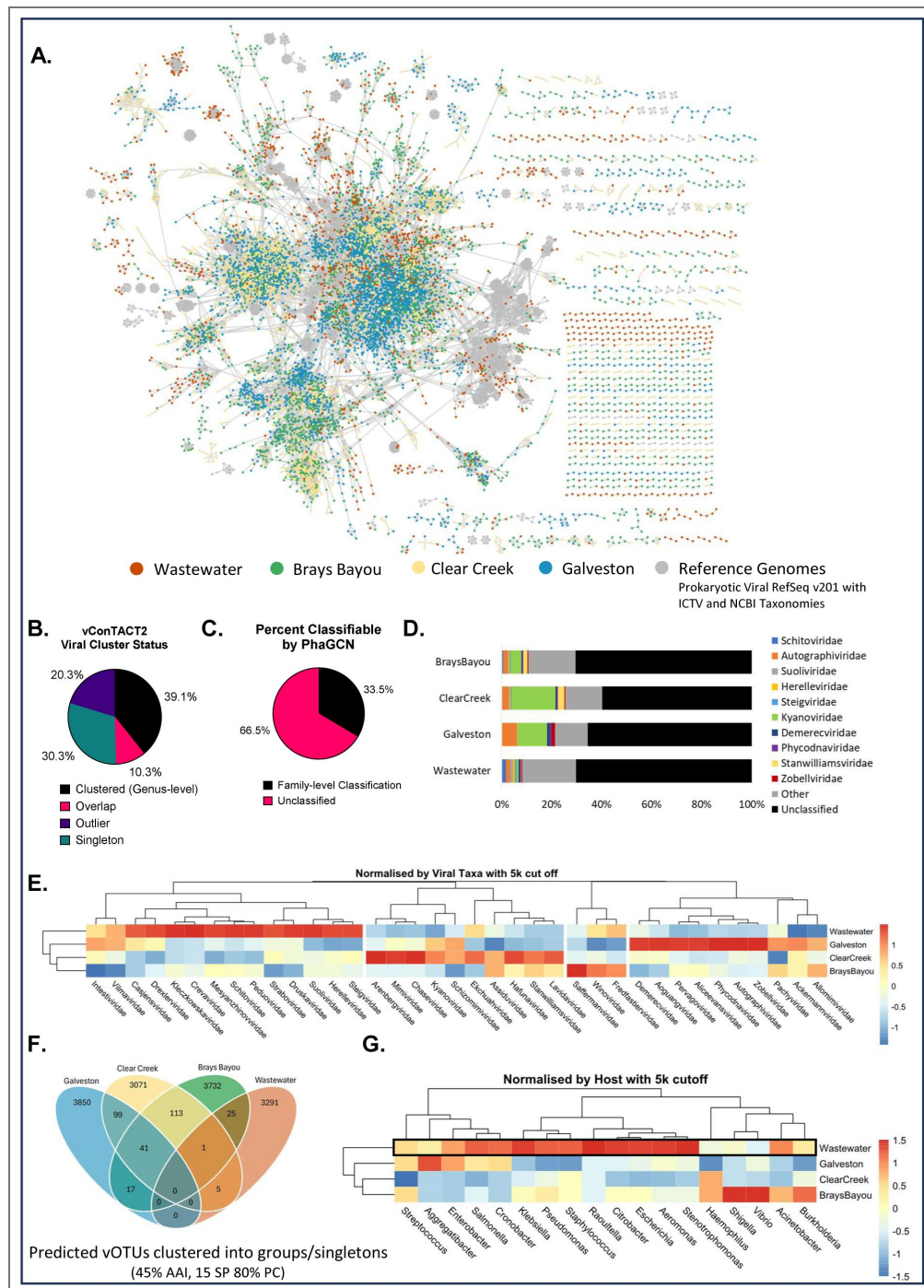


Figure 4. Taxonomic classification of viral contigs from metagenomic datasets reveal distinct viral signatures as well as biased presence of viruses of pathogenic hosts between different locations.

A) vConTACT2 gene-sharing network of 16,198 viral operational taxonomic unit (vOTUs) (≥ 5 kb) from West University WWTP, Brays Bayou, Clear Creek, and Galveston. Each dot (node) represents a vOTU, and each line (edge) represents the similarity between each genome. Additionally, 3,508 reference genomes from the Prokaryotic Viral RefSeq v201 database are shown in grey. **B)** Pie-chart of viral cluster status of vOTUs from vConTact2. **C)** Pie-chart of classified (family-level) vs unclassified vOTU from PhaGCN. **D)** Bar graph of family-level classification (PhaGCN counts) of topmost abundant genera across the samples. Remaining genera are grouped together as "others" in grey. Unclassified vOTUs are in black. **E)** Heat map with z-score normalized by viral taxa to show distinct viral signatures between sites (pheatmap separated each cluster). **F)** Venn-diagram of shared vOTUs (cluster mode: amino-acid identity (AAI) 45%, protein coverage (PC) 80%) depict number of unique vOTUs present at each site and shared between sites. **G)** Heat map with z-score normalized by host reveal the relevance of sampling specific sites for phage hunting.

Kyanoviridae (green) and *Autographiviridae* (orange) made up the largest identifiable taxa in Clear Creek and Galveston (Fig. 4D). This agrees closely with previous findings that *Kyanoviridae*, containing T4-like viruses that infect cyanobacteria, and *Autographiviridae*, T7-like viruses, have been found in deep-sea viromes.⁴⁴ Due to current limitations in viromics, a large portion of our metagenomic dataset remain unknown. Collectively, between 60.9% (outlier/singles from vCONTACT2) and 66.5% (unclassified from PhaGCN) of vOTUs are in this category. A substantial fraction is considered viral dark matter, but we used the remaining identifiable sequences to provide some biological context. This enable us to validate our sampling strategies and compare samples and biomes from another. We used counts from PhaGCN classifications to generate heatmaps normalized by viral taxa to see distinct viral clusters from each biome type. There are four viral signatures present, separating wastewater, seawater, freshwater and brackish water (Fig. 4E). We further clustered viruses from each biome together to identify overlaps and visualized them with Venn diagrams (Fig. 4F). We saw that overlaps of viral genomes reflected the physical environment; for example, the brackish water of Clear Creek overlapped between seawater and freshwater - representing its status as an admixture of the two. Next, we used CHERRY for host prediction and made heatmaps normalized by hosts. A large majority of phages with pathogenic hosts were found in wastewater virome (Fig. 4G, black box).

To increase phage diversity and minimize the emergence of phage resistance, curating a large, genetically diverse phage library is advantageous. Such resource allows researchers and clinicians to select phages tailored to their patient needs, including targeting specific phage receptors, maximizing host coverage, designing phage cocktails, or reducing dependence on any single phage.⁴⁵ We compared *Pseudomonas*- and *Escherichia*-infecting viruses from freshwater and wastewater. Phylogenetic analysis of *Pseudomonas*-infecting viruses showed significant increases in diversity when combining vOTUs from the two biomes (Fig. 5A). Counting and clustering the unique genomes, there were no sequences that were in common between sewage and freshwater (Fig. 5B). Analyzing the *Escherichia*-infecting viruses, phylogenetic diversity was greatly increased by including freshwater phages, and the two biomes had no sequences in common (Fig. 5C,D). This shows the benefit of creating a phage library using Φ HD from multiple biomes.

Resistant phage library (R Φ -Library)

Last-line antibiotics serve as a critical safety net for treating life-threatening infections by multi-drug-resistant bacteria. Their use is restricted to preserve efficacy and slow the development of drug resistance. In a similar manner, the R Φ -library was conceptualized as a last-line phage library to combat extremely phage resistance organisms (X Φ ROs). To test our approach of ge Φ mapping and Φ HD and generate a therapeutic phage library for X Φ ROs that posed clinical challenges, we used 17 X Φ ROs from the TAIL Φ R collection. These failed all conventional phage discovery methods (Fig. 6A) including sewage spotting and Appelmans protocol (for host range expansion). As screening Φ HD concentrates allows searching far larger volumes (Fig. 6B), we immediately found therapeutic phage candidates for all 17 X Φ ROs using freshwater (Brays Bayou) and sewage (West University) sources (Fig. 6C). We purified (Fig. 6D) and characterized 36 new phages by genomic analysis and TEM (Fig. 6D,E, Supp. Table 9). This collection of phages is what we termed the R Φ -library. Compared to TAILOR's sequenced phages, proteomic dendrograms (Fig. 6E) showed phages with both subtle and high genetic variation, including new genera exemplified by MYC30C1 and MYC30C2 against *Enterococcus* (Supp. Fig. 5, 6). Our technique also isolated the jumbo phage UCS29C1 against *E. coli*, which was closely related to the *Klebsiella* jumbo phage Miami, showing that even large phages can survive the process intact (Supp. Fig. 5, 6).⁴⁶ These successes make GE Φ MAPPING and Φ HD sampling excellent additions to TAIL Φ R's phage discovery pipeline, allowing for isolation of novel phages with greater precision and efficiency (Fig. 6F).

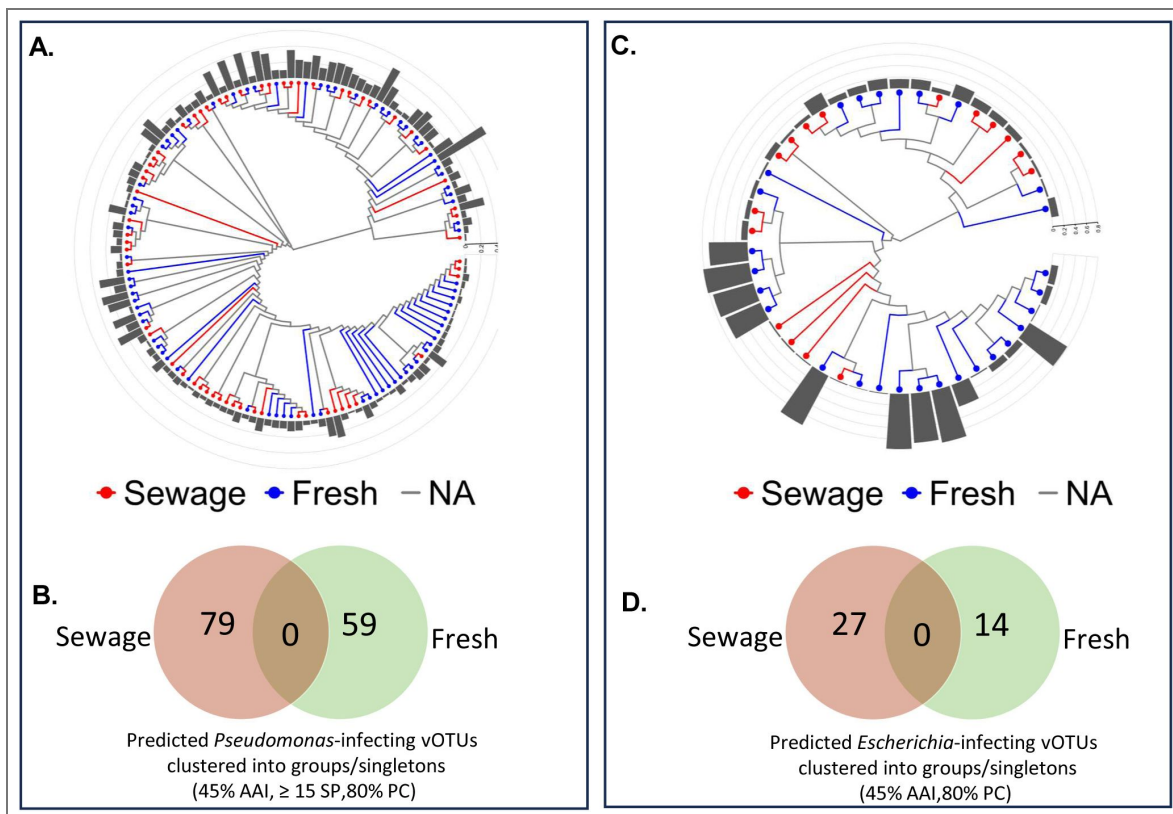


Figure 5. Relevance of multi-site sampling for increased diversity and collection of unique *Pseudomonas* and *Escherichia* phages.

A) Phylogenetic tree of all vOTUs predicted to have *Pseudomonas* host from West University WWTP wastewater and Brays Bayou freshwater metagenomic dataset. Branch-length is non-scaled. **B)** Venn-diagram of vOTUs-grouping (cluster mode: AAI 45%, 15 shared protein (SP), PC 80%) show no common vOTUs shared between sites. **C)** Phylogenetic tree of all vOTUs predicted to have *Escherichia* host from wastewater and Brays Bayou metagenomic dataset. Branch-length is non-scaled. **D)** Venn-diagram of vOTUs-grouping (cluster mode: AAI 45%, 15 shared protein (SP), PC 80%) show no common vOTUs shared between sites. Bar graphs (A, B) represents the genomic similarities between a contig and its closest neighbor based on genome-wide sequence similarities computed by tBLASTx from VipTree.

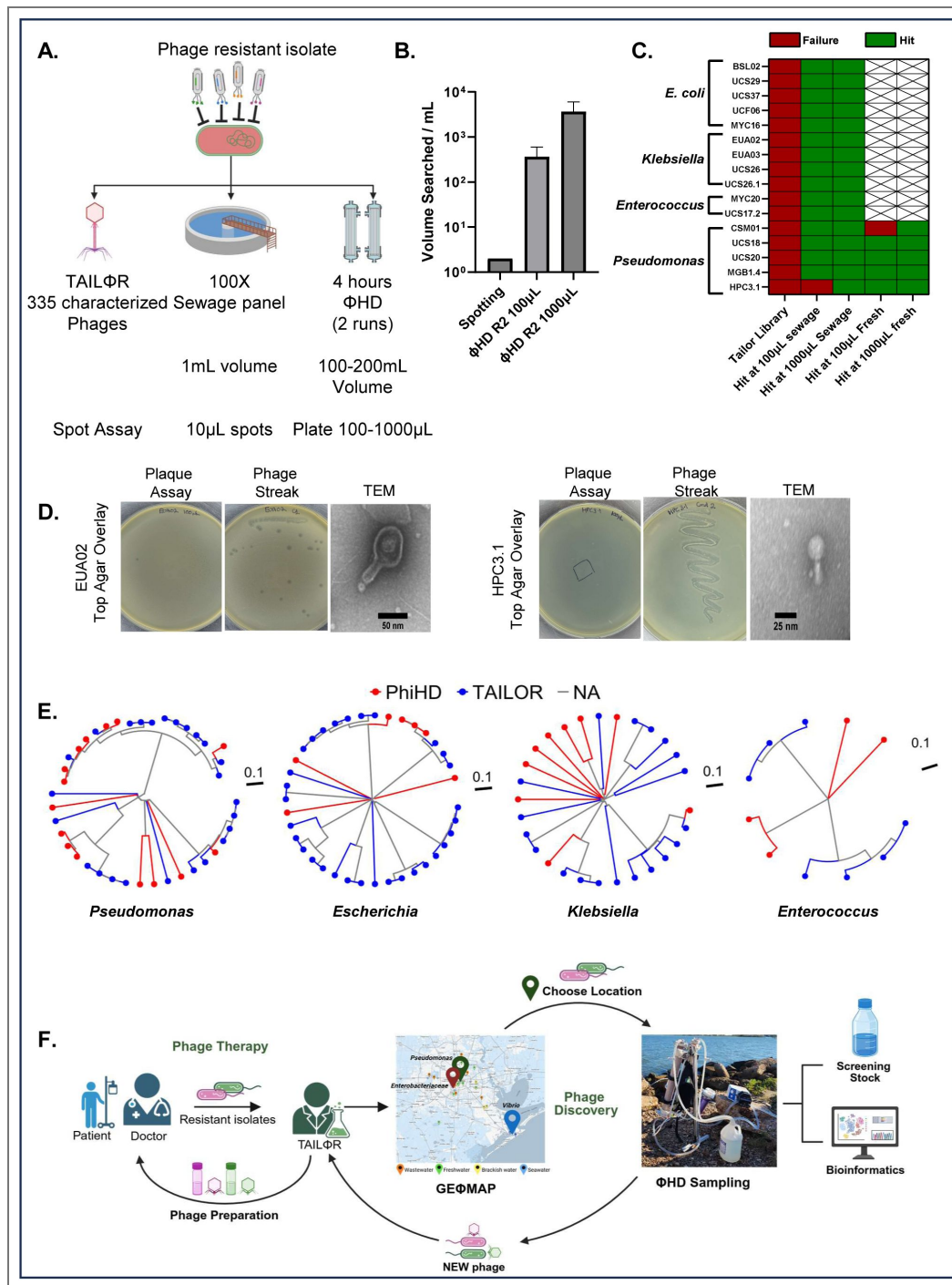


Figure 6. ΦHD increases search volumes and allowed efficient discovery of phages for phage-resistant bacterial pathogens.

A) Schematic of ΦHD compared to standard identification and discovery of therapeutic phages, **B)** Effective volumes that can be searched comparing standard spotting of concentrates with high-concentration, high volume ΦHD retentates (n=2 retentates with differing final concentrations), **C)** Heat map showing success/failure of ΦHD compared to the TAILΦR library in finding therapeutic phage candidates. **D)** Plaque assays (left column) on phage-resistant strains EUA02 (*K. pneumoniae*) and HPC3.1 (*P. aeruginosa*). Individual plaques (middle columns) were streaked and isolated from plaque assays. TEM images (right columns) of select isolated phages infecting EUA02 and HPC3.1. Representative images shown. **E)** Dendrograms of all isolated phages for *Pseudomonas*, *Escherichia*, *Klebsiella*, *Enterococcus*. Phages from ΦHD are in red, and phages from the TAILΦR library in blue. **F)** Diagram of the implementation of ΦHD and geΦmapping to the phage discovery pipeline. Created with BioRender.com.

Discussion

Our study leverages the billion year-long evolutionary war between phages and their hosts, focusing on discovering and harnessing natural phages. We report: i) the design, construction, and successful use of a high-throughput phage capture device (Φ HD); ii) the use of this device to identify regional sources of both bacteria and phages (ge Φ mapping); iii) the use of ge Φ mapping coupled with Φ HD sampling to identify novel phages that target phage-resistant strains, when all other phage matching solutions failed (R Φ -Library); and iv) the characterization of new metagenomes. Collectively, this work provides a rich resource for therapeutic phages against resistant ESKAPEs (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter spp.*) and a rich gene library for new biological discovery (Fig. 6F [↗](#)).

Wastewater, areas of high human activity and dense urban development are particularly rich in phages capable of targeting antibiotic-resistant pathogens.^{14,16,47–49} This is largely due to the elevated microbial loads in these settings, which naturally support a higher abundance and diversity of phages.^{50,51} Human activity imposes evolutionary pressure on bacteria, driving the emergence of antibiotic-resistant strains—and, by extension, their phage counterparts.⁵² To strategically identify optimal sampling locations, we employed ge Φ mapping to pinpoint environments with high pathogen loads (Fig. 2E [↗](#), 3C [↗](#)). We identified Brays Bayou as an excellent locale for anti-*Pseudomonas* and other phages, with agreement between genomic mapping and phenotypic phage surveys. We extended our approach to sewage, not only localizing to a single WWTP for phage hunting but to a specific section of that plant (Fig. 3 A–F [↗](#)).

This targeting enabled the use of Φ HD which produced phage-rich retentates with phage yields increasing by 100–3500-fold each sample in comparison to unprocessed input (Fig. 2F [↗](#), 3H [↗](#), Supp. Fig. 2G [↗](#)). These retentates were used to target X Φ ROs in clinical cases resulting in not only the generation of the R Φ -library (Fig. 6C [↗](#)) and discovery new genera of phages (Supp. Fig. 5, 6 [↗](#)) but also treatment in patient cases. Currently, 1 patient has received 2 phages that necessitated the Φ HD approach, and 6 more patient cases are in progress that required Φ HD-derived phages (not published). This illustrates the critical need of phage discovery being addressed by the ge Φ mapping and Φ HD approach.

Our study is limited by multiple factors. Our current ge Φ map is limited to the Gulf Coast area around Houston, TX, USA and thus is of limited utility to others. However, the overall approach of localizing phage reservoirs is generalizable. With regards to the geomap-guided phage hunt, we only compared one low host-availability control site (Clear Creek) with two high host-availability sites (wastewater and Brays Bayou). More comparisons, especially incorporating even more sites with varied host availability and including additional hosts to screen, would strengthen the claim. Beyond our control, we were limited by weather conditions, as temperature and rainfall can drastically alter both microbial and viral composition. In turn, this limited our sampling time to days with consistent weather (i.e., dry, warm days) for accurate comparisons. In the field, it is technically challenging to remove all phages and bacteria from the hollow fiber units at the end of a sampling run. We employ simple backflushing through the retentate port which increases phage yield by 25–50%. More phages could be stuck in the hollow fibers. Given the amount of particulate removed later during cleaning, we are certain recovery percentages can be further increased with additional backflushing. While the Φ HD system is portable, the total assembly is large and requires at least two people to carry and operate. We are currently developing a smaller, portable version for single person use to address this. This would allow for researchers to sample in more remote locations and require less manpower. Finally, our approach is limited to aquatic biomes, but it is well known that the soils of forests and other areas (animal waste, sand, and so on) harbor huge phage biodiversity. We are currently working on ways to address this for both viral ecology and clinical phage discovery.

Collectively, we used a bioprospecting approach to localize therapeutic phages and harvest them *en masse* with Φ HD. This resulted in large metagenomic data sets informing the viral ecology of the Gulf Coast, critical therapeutic phages for challenging clinical cases and tools for others to use

in phage discovery.

Methods

Sample acquisition and preparation for 16S mapping

4L freshwater and seawater samples from 15 locations were collected in containers cleaned with 10% bleach and copiously pre-rinsed in ddH₂O. Locations and weather were precisely recorded (Supp. Table 10⁵³). Maps were made with EasyMapMaker.⁵³ Samples were collected in triplicates with the exception to Brays Bayou @ HEB. Based on 16S mapping and data not shown here, this location is biologically equivalent to Brays Bayou @ Medical Center due to its short distance and lack of inflow between them. For sampling purposes, we used this location due to its easier access to the bayou and safety reasons. Samples were transported back to the laboratory within hours of gathering samples and stored at 4°C until further processing next day. We centrifuged approximately 400 mL of each sample at 10,000xg for 20 minutes to pellet bacteria and debris. 250 mg was measured out in pre-weighed tubes. In some samples, where there were insufficient pellets (e.g. Buffalo @ Santa Ana Capture Site, Galveston, West Bay, Sims Bayou, Halls @ Keith Weiss Park), we centrifuged an additional 400mL of water until 250 mg was met. Sewage samples were obtained from the influent tank of various wastewater treatment plants by plant workers and transported to BCM and handled in a similar fashion. We centrifuged 50 mL wastewater influent at 10,000xg for 20 minutes. 250 mg of the pellet was measured in pre-weighed tubes.

DNA preparation and 16S processing

DNA was extracted using the DNeasy PowerSoil Kit (Qiagen, Germany) according to the manufacturer's instructions. Quality was evaluated by A260/280 ratio using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher, USA). Samples with a A260/280 ratio outside of the 1.8-2.0 range were further processed using a Monarch Genomic DNA Purification kit (NEB, USA) using the "clean-up" protocol in the manufacturer's instructions. Samples were reprocessed from the start if DNA concentrations were insufficient. Samples were sent for sequencing to Novogene. The V4 region of the 16S gene was amplified using the 515F (5'-GTGCCAGCMGCCGCGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3') primers^{54,55} and then sequenced on a Novaseq6000 and 100K raw reads collected (Novogene, China). We included negative controls with each sequencing batch, consisting of the PowerSoil kit without sample to generate a "kitome" to assess the laboratory and kit background. For these control samples, 16S rRNA amplification and sequencing was completed despite lack of high-quality DNA inputs. Raw sequences were uploaded to SRA (BioProject #: PRJNA1309115).

16S data analysis

Raw reads were processed in mothur (v1.48.0) guided by the MiSeq SOP using SILVA (v132) reference files and seed.⁵⁶⁻⁵⁹ Owing to the size, number of samples, and relative lack of computational resources we used the cluster.split command with taxlevel=4 to cluster the sequences into OTUs. We generated rarefaction curves using default parameters and calculated alpha-diversity statistics in mothur, subsampling 50,000 random sequences from each group. Beta diversity analysis and PCoA was likewise performed in mothur using Bray-Curtis distance matrices and visualized in Graphpad Prism (v.10.1.2, GraphPad Software, USA). Using R, we performed AMOVA to confirm our β -diversity comparisons. To analyze and visualize the identified taxa in each sample we generated Krona charts on a Galaxy web server.^{60,61} We generated heatmaps of the relative amounts of specified taxa by calculating z-scores in R.⁶² We calculated z-scores both by site and by taxa, and displayed both using pheatmap.⁶³

ΦHD environmental sampling

A filtration/concentration device, ΦHD, was designed to filter microbes and phages from water. The filtration train starts with a 297 μ m stainless steel mesh hose filter (3.6" in height and 2" in width) attached to silicone peristaltic tubing and goes into a water source (the feed line). The hose

filter excludes trash, plant matter, and fish from the system. The tube connects to a Countertop Filtration System equipped sequentially with a 25 μm blown polypropylene cartridge and a 5 μm spun polypropylene cartridges. The filters remove more debris (soil, sand, etc.) and larger microorganisms like bacteria. Tubing then connects to the feed port of a 100 kDa hollow fiber unit (modified polyether sulfone (mPES), 65 cm, KrosFlo Hollow Fiber Filter Module). Anything smaller than 100 kDa (viruses, smaller bacteria, etc.) is filtered by tangential flow filtration (TFF) and filtered through the permeate ports. The retentate port ejects filtered water into a (pre-cleaned with 10% bleach and diH_2O -rinsed) 4L reservoir. Stop valves are connected to the retentate line and bottom permeate line to control flow rates and fill the jacket of the unit. We use one head of a dual-headed peristaltic pump to propel water through the filtration train. The pump is powered with a travel-size battery station, Jackery Explorer 300, that was charged with solar panels.

Water was filtered and concentrated into a single reservoir, and a second 100 kDa hollow fiber unit was used to further concentrate the water into smaller volumes. The second head of the pump continuously cycled retentates from reservoir to unit. We used a custom-built aluminum stand that allows easy field use of the hollow fiber units. The concentrated product was expected to be 500 mL to 1 L in volume. ΦHD sampling in-field is the first stage of processing (Supp. Fig 7A [↗](#)). Samples were brought back to the lab in 1-2 hours. We centrifuged the samples, generally 1-2L, at 10,000xg for 20 minutes and filtered the samples with 0.22 μm vacuum-filters (Steritop filters (PES). Samples were stored at 4°C until second stage processing in lab (next day). A smaller version of the 100 kDa hollow fiber unit was used to achieve volumes of 50-200 mL per sample (mPES, 20 cm, MidiKros Hollow Fiber Filter Module). Retentate flow rates are controlled by Luer lock stoppers on the retentate line and bottom permeate line. Concentrations of each sample were recorded accordingly after secondary concentration (Supp. Table 5 [↗](#)).

Cleaning-in-place (CIP) protocol of hollow fiber units (TFF units)

After each ΦHD run, we cleaned the TFF units to regenerate them. For 5-10 mins, water was flushed through the permeate ports on both ends of the TFF unit. This reverse flow removes the majority of biological build-up in the units. The outer jacket and hollow fibers were then filled with 0.5M HCl and left overnight to incubate at room temperature. The next day, we rinsed the units with water then soaked outer jacket and hollow fibers in 0.5M NaOH overnight. Once rinsed with diH_2O , the units are ready to be used again. For more polluted waters (such as wastewater), we repeated the CIP protocol. We've successfully used the units over 10 times for this study (Supp. Table 5 [↗](#)).

ΦHD wastewater sampling

Wastewater has a high burden of flotsam (toilet paper, wet wipes and feminine hygiene production, etc.) as well as smaller particulates (Supp. Fig. 7B [↗](#)). We pre-filtered wastewater and used the pre-filtered input as the feed to ΦHD . Two silicone peristaltic tubes drew wastewater 1-2 meters from the influent tank of West University WWTP and into two large tubs, pre-cleaned with 70% ethanol and Clorox wipes. This helped avoid gravitation issues such as slower flow rates and battery tripping. Wastewater was allowed to settle for 30 minutes. This allowed larger, denser material to sediment to the bottom. Two tubes with styrofoam floaters and barbed hose filters propelled water through two spin down filters – containing 150 and 63 μm mesh cartridges. The floaters were essential in avoiding the settled debris at the bottom, and the larger filters helped remove more debris. Wastewater was stored in two pre-cleaned tubs; this became the input for ΦHD where we processed the water as described above. For wastewater processing, an additional valve on the feed line is needed to slow down flow rate. Similarly, we centrifuged the first stage concentrate (1-2L) at 10,000g for 20mins and filtered the samples with 0.22 μm vacuum filters. second stage processing lowered the volume to 150-200 mL.

Bacterial growth and plaque assays

A table of strains used in the study can be found in [Supp. Table 11](#). From a single colony, bacterial cultures were inoculated in Luria Broth (LB, Sigma-Aldrich) and grown overnight at 37°C with shaking at 180 rpm. For plaque assays, overnight cultures were suspended in 0.75% LB top agar with 10-1000 µL of ΦHD concentrate and poured onto 1.5% LB plates to solidify. Plates were incubated overnight at 37°C. Plaques are counted next day. Phage titers were visualized with Graphpad Prism (v.10.6.1, GraphPad Software, USA). For phage streaks, a single plaque on agar plates was picked and resuspended in 25 µL of phosphate buffer saline. Bacterial cultures were suspended on 0.75% LB top agar and poured onto LB plates. Before the top agar solidifies, the resuspended plaque was streaked with a pipette. To isolate individual phages from a plaque, the phage was continuously streaked out until plaque morphologies are consistent (at least twice).

DNA preparation and shotgun shallow metagenomic sequencing

11.5 mL of second stage concentrates from ΦHD ([Supp. Table 5](#)) were centrifuged with a SW-41Ti rotor at 37.5k rpm for 2 hours to pellet virus-like particles (VLPs). Supernatant was removed, and the pelleted viruses are used for DNA extraction using the DNeasy PowerSoil kit (Qiagen, Germany) according to manufacturer's protocols. The extracted DNA was sent for PCR-free library preparation followed by shallow shotgun metagenomic sequencing on the NovaSeq X Plus system platform using the PE150 strategy (Novogene, Beijing). This resulted in >2G of raw data per sample with Q30>85. Raw sequences were uploaded to SRA (BioProject #: PRJNA1308632).

Viral metagenomic analysis

Raw data was uploaded into KBASE⁶⁴ and examined with FastQC (0.12.1)⁶⁵ then trimmed with Trimmomatic (v0.36)⁶⁶ with default settings (sliding window size 4, sliding minimum quality 15, post tail crop length null, head crop length 0, leading minimum quality 3, trailing minimum quality 3, minimum read length 36).⁶⁶ Post-trimmed reads were assessed with FastQC (0.12.1). Reads were assembled into contigs with metaSPAdes (v3.15.3); minimum contig length was $300 \leq 2000$ bp and k-mer sizes were 21,33,55,77,99, and 127.⁶⁷ Contigs were uploaded into CyVerse for further analysis with VIBRANT.^{68,69} VIBRANT was used to identify and extract viral contigs from the dataset; default settings were used (length of bp was 1000, number of ORFs per scaffold was 4). Viral contigs from individual samples were clustered into viral operational taxonomic unit (vOTU) using average nucleotide identity (ANI)-based clustering (95% ANI, 85% alignment coverage) on PhaBOX.⁷⁰ We used CheckV⁷¹ to assess quality and completeness of vOTUs. vOTUs were annotated with PROKKA (v1.14.5)⁷² on KBASE, and gene-sharing networks were made with vConTACT2 (v0.9.19).⁷³ Settings were set to default, and the reference database used for cluster taxonomy was NCBI Bacterial and Archaeal Viral RefSeq V201. We visualized the gene-networks with Cytoscape.⁷⁴ vOTUs were further taxonomically classified with PhaGCN⁷⁵ (minimum length ≥ 5 kb, 75% amino acid identity (AAI), shared protein 15, and 80% protein coverage) and predicted host of each vOTUs with CHERRY⁷⁶ (minimum length ≥ 5 kb, 75% amino acid identity (AAI), shared protein 15, and 80% protein coverage, 90% CRISPRs identity). Again, we generated heatmaps of the relative amounts of identified taxa or predicted hosts by calculating z-scores and displayed them using heatmap in RStudio.^{62,63}

vOTUs with predicted hosts in the genera *Pseudomonas* and *Escherichia* were manually extracted from the datasets of wastewater (West University WWTP) and freshwater (Brays Bayou). We clustered vOTUs with AAI-based clustering mode on PhaBOX (45% AAI, 15 shared protein, and 80% protein coverage). We used VIPTree⁷⁷ phylogenetic analysis then visualized the resulting Newick trees using ggtree⁷⁸ in RStudio.

Statistical analysis of metagenomic sequences

Raw data was uploaded into KBASE and examined with FastQC (0.12.1) then trimmed with Trimmomatic (v0.36) with default settings (sliding window size 4, sliding minimum quality 15, post tail crop length null, head crop length 0, leading minimum quality 3, trailing minimum quality 3,

minimum read length 36). Post-trimmed reads were assessed with FastQC (0.12.1).

We used MetaPop to analyze viral metagenomic sequence data at the interpopulation (macrodiversity) level.⁷⁹ We aligned and mapped filtered, trimmed reads of each biome to our reference genome with bowtie2 (v2.5.4+galaxy0). The 16,198 vOTUs (95% ANI, 85% alignment coverage) from wastewater, seawater, brackish, and freshwater were combined and used as our reference genome. Generated BAM output files from bowtie2 and the reference genome were used as input for MetaPop. MetaPop's macrodiversity analysis includes raw population abundance, normalized population abundance, and α -diversity calculations. The following analyses were conducted in RStudio using tidyverse, vegan, and pheatmap. From normalized population abundance tables, we calculated z-scores and derived heatmaps to display feature-level patterns. We calculated β -diversity using Bray-Curtis dissimilarity, then we performed PCoA with the ordination of the Bray-Curtis matrix. To assess robustness, 2,000 features were randomly subsampled and analysis repeated across 1,000 bootstrap iterations. Resulting ordinations were aligned to a reference with Procrustes alignment. Mean coordinates and standard deviations were calculated for each sample, and scatter plots were generated.

Generation of R Φ collection

We obtained clinical isolates from the collection of TAIL Φ R Labs at Baylor College of Medicine (BCM). We selected isolates for which no phage had been isolated despite exhaustive searches in sewage spotting and testing of the TAIL Φ R phage library. Study of these clinical isolates was approved by BCM's Institutional Review Board (IRB). We grew each isolate overnight in LB media, mixed 100 μ L of this overnight culture with 3 mL of 0.7% top agar and either 100 μ L or 1 mL of environmental or sewage Φ HD concentrate as indicated. Plates were incubated overnight at 37°C. Next day, plaques that we considered as lytic in appearance (non-cloudy, cleared zones on bacterial lawns) were picked with a pipette tip and streaked in top agar with the same clinical strain prepared in an identical manner. Plaques were streaked 3 times to ensure purity. We endeavored to pick 2-3 different plaque morphologies per clinical isolate for each source (environment or sewage). The purified plaques were picked and mixed in 100 μ L of phage buffer (10 mM Tris pH 8.0, 100 mM NaCl, 10mM MgSO₄). We then made plate lysates by mixing 50 μ L of this phage-laden material with 100 μ L bacterial overnight of the indicated strain and 3 mL 0.7% top agar and incubating overnight. These were harvested by scraping the top agar from the plates, rinsing the plate surface with 3 mL phage buffer and centrifuging this mixture at 2200xg for 10 minutes and removing residual bacteria with a 0.22 μ m PVDF filter (Merck Millipore, Ireland). These phages were titered using standard methods on 0.75% top agar on the indicated strain. Anti-HPC3.1 phages underwent plate lysate production in a laboratory strain PAO1 owing to the amount of mucus produced, and were subsequently titered on the indicated clinical strain. Any other phages found in this study was noted in [Supp. Table 12](#) [↗](#).

Genetic analysis of the R Φ collection

We employed various methods of DNA preparation for the phages. For samples with adequate titers, we used the EZNA Universal Pathogen Kit (Omega Bio-Tek, USA) according to the manufacturer's instructions. Phage strains with lower titers or that failed to produce adequate DNA using the EZNA kit were processed by generation of 9 mL of plate lysate (as above) and centrifugation with a Sw-41Ti rotor at 37.5k rpm for 2 hours to pellet virus-like particles (VLPs). This pellet was then processed using a Qiagen Power Soil Kit using the manufacturer's instructions. For DNA samples with ratios outside of 1.8-2.0, we further purified them with the NEB Genomic DNA Purification kit using the manufacturer's "clean-up" protocol. Whole genome sequencing using standard protocols was done by BCM's CMMR core. We processed short reads data using FastQC (v0.12.1)⁶⁵ to assess data quality, removed low quality sequences (Q>30 cut off) and removed the adapters using Trimmomatic (v0.39)⁶⁶ with default settings. We verified removal of adapters and quality cutoffs with another run of FastQC. These clean reads were then assembled using SPAdes (v3.15.3) (DNA source standard, minimum contig length 200 $\leq$ 500)⁸⁰ and annotated using RASTtk (v1.073) (default setting). We identified candidate viral sequences from

these assemblies using ⁸¹VirSorter2 (discarding sequences with <2 hallmark genes, requiring hallmark genes on all sequences, only output high confidence viral sequences, and requiring viral genes to be annotated).⁸² Sequences with high dsDNA phage score, >4-6 hallmark genes, relatively low cellular gene fraction and coverage suggestive of high copy number above the bacterial host were examined manually and BLASTn⁸³ was used to determine phage or not. Confirmed phage sequences were then re-annotated with RASTtk (v1.073) to provide final phage genomes for the RΦ library. We compared TAILOR sequences (pre-existing data) with the sequences of purified phages from RΦ library. We used VIPTree⁷⁷ for a proteomic dendrogram and then visualized the resulting Newick trees using ggtree⁷⁸ in RStudio. Several Klebsiella phages that were isolated on related host strains were identical and so we removed these duplicates prior to visualization. All phage genomes were re-annotated with Pharokka v 1.9.0.⁸⁴ Coding sequences were predicted with PHANOTATE⁸⁵ tRNAs were predicted with tRNAscan-SE 2.0.⁸⁶ tmRNAs were predicted with Aragorn.⁸⁷ CRISPRs were predicted with CRT.⁸⁸ Functional annotations were generated by matching each CDS to the PHROGs⁸⁹, VFDB⁹⁰, and CARD⁹¹ databases using MMseqs2^{92–94} and PyHMMER.⁹⁵ Phage genomes were rotated with Dnaapler.⁹⁶ Contigs were matched to their closest hit in the INPHARED⁹⁷ database using mash.⁹⁸ GenBank accession numbers for phage genomes are on [Supp. Table 9](#) [and](#) [Supp. Table 12](#) [.](#)

Electron microscopy and figures

Various environmental, sewage and purified phage samples were imaged at BCM's Cryo-EM Advanced Technology Core Facility. Samples were deposited on grids (Quantifoil 2/2 200 Cu + 2-nm ThinC; Quantifoil Micro Tools GmbH, Jena, Germany) and negatively stained with 2% uranyl acetate using standard techniques. Images were collected with either a JEOL 1230 electron microscope (JEOL, Japan) at 80 kV equipped with a 4,000 by 4,000 Gatan Ultrascan charge-coupled device camera (Gatan, Ametek) or a JEOL 2100 (200 kV) electron microscope equipped with a Gatan 4k x 4k CCD. Images were collected at numerous magnifications, and representative examples were selected and viewed in NIH ImageJ.

Data availability

Raw shotgun metagenomic sequences were uploaded to NCBI SRA (Sequence Read Archive) under BioProject PRJNA1308632. 16S rRNA amplicon sequences were uploaded to NCBI SRA under BioProject PRJNA1309115. Genome sequences generated in this study have been deposited in the National Center for Biotechnology Information (NCBI) GenBank database, with accession numbers provided in the Supplementary Information.

Acknowledgements

This work was supported by the National Institutes of Health (grant number 5 U19 AI157981) with additional funding from the Levy-Longenbaugh Fund and Kleberg Foundation. PN received additional support from T32 AI0554413 fellowship, K08 AI173452-01A1, and BCM Intramural funds. We thank Alleigh Nicholls, Tai Gip, Christian Maresso, Dylan Chirman, Ling Qiu, Ellen Vaughan, John Taylor and Rachel Lahowetz for their assistance on in-field sampling. CryoEM data was collected at the Baylor College of Medicine CryoEM ATC, which includes equipment purchased under support of CPRIT Core Facility Award RP190602.

Additional files

[Supplementary Figures and Tables](#) [.](#)

Additional information

Funding

Funder	Grant reference number	Author
National Institute of Health Sciences (NIHS)	5 U19 AI157981	Camilla Do
		Anthony W Maresso
		Justin R Clark
		Austen Lee
		Terwilliger
		Paul Nicholls
		James D Chang
		Keiko Christine Salazar
		Paul Ruchhoeft
National Institute of Health Sciences (NIHS)	T32 AI0554413 Fellowship	Paul Nicholls
National Institute of Health Sciences (NIHS)	K08 AI173452-01A1	Paul Nicholls
Baylor College of Medicine (BCM)	Intramural Funds	Paul Nicholls
Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation (Kleberg Foundation)		Camilla Do
		Anthony W Maresso
		Keiko Christine Salazar
		Paul Nicholls
		Paul Ruchhoeft
		Austen Lee
		Terwilliger
		James D Chang
		Justin R Clark
Levy-Longenbaugh Fund		James D Chang
		Anthony W Maresso
		Paul Nicholls
		Austen Lee
		Terwilliger
		Paul Ruchhoeft
		Keiko Christine Salazar
		Justin R Clark
		Camilla Do

Author ORCID iDs

Camilla Do:  <https://orcid.org/0000-0002-8798-908X>

Anthony W Maresso:  <https://orcid.org/0000-0002-4452-3490>

References

- 1 **Naghavi M**, et al. (2024) Global burden of bacterial antimicrobial resistance 1990-2021: a systematic analysis with forecasts to 2050. *The Lancet* **404**:1199-1226 [https://doi.org/10.1016/S0140-6736\(24\)01867-1](https://doi.org/10.1016/S0140-6736(24)01867-1) | PubMed
- 2 **Pirnay J.-P**, et al. (2024) Personalized bacteriophage therapy outcomes for 100 consecutive cases: a multicentre, multinational, retrospective observational study. *Nature Microbiology* **9**:1434-1453 <https://doi.org/10.1038/s41564-024-01705-x> | PubMed

- 3 Green SI, et al. (2023) A Retrospective, Observational Study of 12 Cases of Expanded-Access Customized Phage Therapy: Production, Characteristics, and Clinical Outcomes. *Clinical Infectious Diseases* **77**:1079-1091 <https://doi.org/10.1093/cid/ciad335> | PubMed
- 4 Hutinet G, et al. (2019) 7-Deazaguanine modifications protect phage DNA from host restriction systems. *Nature Communications* **10**:5442 <https://doi.org/10.1038/s41467-019-13384-y> | PubMed
- 5 Katz MA, et al. (2024) Diverse viral cas genes antagonize CRISPR immunity. *Nature* **634**:677-683 <https://doi.org/10.1038/s41586-024-07923-x> | PubMed
- 6 Garneau JE, et al. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**:67-71 <https://doi.org/10.1038/nature09523> | PubMed
- 7 He L, et al. (2024) Tail assembly interference is a common strategy in bacterial antiviral defenses. *Nature Communications* **15**:7539 <https://doi.org/10.1038/s41467-024-51915-4> | PubMed
- 8 Oechslin F (2018) Resistance Development to Bacteriophages Occurring during Bacteriophage Therapy. *Viruses* **10** <https://doi.org/10.3390/v10070351> | PubMed
- 9 Terwilliger AL, et al. (2020) Tailored Antibacterials and Innovative Laboratories for Phage (Φ) Research: Personalized Infectious Disease Medicine for the Most Vulnerable At-Risk Patients. *Phage* **1**:66-74 <https://doi.org/10.1089/phage.2020.0007> | PubMed
- 10 Rammelkamp CH, Maxon T (1942) Resistance of Staphylococcus aureus to the Action of Penicillin. *Proceedings of the Society for Experimental Biology and Medicine* **51**:386-389 <https://doi.org/10.3181/00379727-51-13986>
- 11 Olsen NS, Hendriksen NB, Hansen LH, Kot W (2020) A New High-Throughput Screening Method for Phages: Enabling Crude Isolation and Fast Identification of Diverse Phages with Therapeutic Potential. *Phage* **1**:137-148 <https://doi.org/10.1089/phage.2020.0016> | PubMed
- 12 Dedrick RM, et al. (2019) Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. *Nature Medicine* **25**:730-733 <https://doi.org/10.1038/s41591-019-0437-z> | PubMed
- 13 Yehl K, et al. (2019) Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. *Cell* **179**:459-469.e459 <https://doi.org/10.1016/j.cell.2019.09.015> | PubMed
- 14 Sada TS, Tessema TS (2024) Isolation and characterization of lytic bacteriophages from various sources in Addis Ababa against antimicrobial-resistant diarrheagenic Escherichia coli strains and evaluation of their therapeutic potential. *BMC Infectious Diseases* **24** <https://doi.org/10.1186/s12879-024-09152-z> | PubMed
- 15 Van Twest R, Kropinski AM, Clokie Martha RJ, Kropinski Andrew M (2009) *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions* Humana Press. pp. 15-21
- 16 Aghaee BL, Mirzaei MK, Alikhani MY, Mojtahedi A (2021) Sewage and sewage-contaminated environments are the most prominent sources to isolate phages against Pseudomonas aeruginosa. *BMC Microbiology* **21** <https://doi.org/10.1186/s12866-021-02197-z> | PubMed
- 17 Kenney PO, Gómez-Duarte OG (2024) Low-volume enrichment method supports high throughput bacteriophage screening and isolation from wastewater. *PLoS One* **19**:e0298833 <https://doi.org/10.1371/journal.pone.0298833> | PubMed
- 18 Muniesa M, Blanch AR, Lucena F, Jofre J (2005) Bacteriophages may bias outcome of bacterial enrichment cultures. *Appl Environ Microbiol* **71**:4269-4275 <https://doi.org/10.1128/aem.71.8.4269-4275.2005> | PubMed
- 19 Roux S, et al. (2016) Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**:689-693 <https://doi.org/10.1038/nature19366> | PubMed
- 20 Gios E, Mosley OE, Hoggard M, Handley KM (2024) High niche specificity and host genetic diversity of groundwater viruses. *The ISME Journal* **18** <https://doi.org/10.1093/ismejo/wrae035> | PubMed
- 21 Xiong L, et al. (2023) Revealing viral diversity in the Napahai plateau wetland based on metagenomics. *Antonie van Leeuwenhoek* **117** <https://doi.org/10.1007/s10482-023-01912-2> | PubMed

- 22 **Chen T**, et al. (2023) Virus–pathogen interactions improve water quality along the Middle Route of the South-to-North Water Diversion Canal. *The ISME Journal* **17**:1719–1732 <https://doi.org/10.1038/s41396-023-01481-2> | PubMed
- 23 **Labbé M**, Girard C, Vincent WF, Culley AI (2020) Extreme Viral Partitioning in a Marine-Derived High Arctic Lake. *mSphere* **5** <https://doi.org/10.1128/msphere.00334-20>
- 24 **Malki K**, et al. (2020) Prokaryotic and Viral Community Composition of Freshwater Springs in Florida, USA. *mBio* **11** <https://doi.org/10.1128/mbio.00436-20>
- 25 **Angly FE**, et al. (2006) The marine viromes of four oceanic regions. *PLoS Biol* **4**:e368 <https://doi.org/10.1371/journal.pbio.0040368> | PubMed
- 26 **Schoenfeld T**, et al. (2008) Assembly of viral metagenomes from yellowstone hot springs. *Appl Environ Microbiol* **74**:4164–4174 <https://doi.org/10.1128/aem.02598-07> | PubMed
- 27 **Liu CG**, Thompson BE, Chang JD, Min L, Maresso AW (2024) Construction and characterization of DNA libraries from cultured phages and environmental viromes. *Applied and Environmental Microbiology* **90**:e01171–01124 <https://doi.org/10.1128/aem.01171-24> | PubMed
- 28 **Heffron J**, McDermid B, Mayer BK (2019) Bacteriophage inactivation as a function of ferrous iron oxidation. *Environmental Science: Water Research & Technology* **5**:1309–1317 <https://doi.org/10.1039/C9EW00190E>
- 29 **Langenfeld K**, Chin K, Roy A, Wigginton K, Duhaime MB (2021) Comparison of ultrafiltration and iron chloride flocculation in the preparation of aquatic viromes from contrasting sample types. *PeerJ* **9**:e11111 <https://doi.org/10.7717/peerj.11111> | PubMed
- 30 **Roux S**, et al. (2019) Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology* **37**:29–37 <https://doi.org/10.1038/nbt.4306> | PubMed
- 31 **Roux S**, Hallam SJ, Woyke T, Sullivan MB (2015) Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**:e08490 <https://doi.org/10.7554/eLife.08490> | PubMed
- 32 **Huss P**, et al. (2025) Engineering bacteriophages through deep mining of metagenomic motifs. *Sci Adv* **11**:eadt6432 <https://doi.org/10.1126/sciadv.adt6432> | PubMed
- 33 **Thurber RV**, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. *Nature Protocols* **4**:470–483 <https://doi.org/10.1038/nprot.2009.10> | PubMed
- 34 **Gowers GF**, et al. (2019) Entirely Off-Grid and Solar-Powered DNA Sequencing of Microbial Communities during an Ice Cap Traverse Expedition. *Genes* **10** <https://doi.org/10.3390/genes10110902> | PubMed
- 35 **Clokier MR**, Millard AD, Letarov AV, Heaphy S (2011) Phages in nature. *Bacteriophage* **1**:31–45 <https://doi.org/10.4161/bact.1.1.14942> | PubMed
- 36 **Chibani-Chennoufi S**, Bruttin A, Dillmann ML, Brüssow H (2004) Phage–host interaction: an ecological perspective. *J Bacteriol* **186**:3677–3686 <https://doi.org/10.1128/jb.186.12.3677-3686.2004> | PubMed
- 37 (no date) US OIL RECOVERY, PASADENA, TX. <https://cumulis.epa.gov/supercpad/cursites/csinfo.cfm?id=0607093>
- 38 **Kelly T** (2024) Hundreds of dead fish found floating in popular metro Houston park, investigation underway. <https://www.click2houston.com/news/local/2024/06/11/hundreds-of-dead-fish-found-floating-in-popular-metro-houston-park-investigation-underway/>
- 39 **Fulton J**, Douglas T, Young & Mark, Clokier Martha RJ, Kropinski Andrew M (2009) *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions* Humana Press. pp. 43–54
- 40 **Brumfield KD**, et al. (2022) Microbiome Analysis for Wastewater Surveillance during COVID-19. *mBio* **13**:e0059122 <https://doi.org/10.1128/mbio.00591-22> | PubMed

- 41 Drulis-Kawa Z, et al. (2011) Isolation and characterisation of KP34—a novel ϕ KMV-like bacteriophage for *Klebsiella pneumoniae*. *Applied Microbiology and Biotechnology* **90**:1333-1345 <https://doi.org/10.1007/s00253-011-3149-y> | PubMed
- 42 Bovio-Winkler P, et al. (2023) Genome-centric metagenomic insights into the role of Chloroflexi in anammox, activated sludge and methanogenic reactors. *BMC Microbiology* **23** <https://doi.org/10.1186/s12866-023-02765-5> | PubMed
- 43 Burton FL, et al. (2013) *Wastewater Engineering: Treatment and Resource Recovery* McGraw-Hill Education.
- 44 Zheng K, et al. (2025) VITAP: a high precision tool for DNA and RNA viral classification based on meta-omic data. *Nature Communications* **16**:2226 <https://doi.org/10.1038/s41467-025-57500-7> | PubMed
- 45 Markwitz P, et al. (2021) Genome-driven elucidation of phage-host interplay and impact of phage resistance evolution on bacterial fitness. *The ISME Journal* **16**:533-542 <https://doi.org/10.1038/s41396-021-01096-5> | PubMed
- 46 Mora D, et al. (2021) Complete Genome Sequence of *Klebsiella pneumoniae* Jumbo Phage Miami. *Microbiol Resour Announc* **10** <https://doi.org/10.1128/mra.01404-20> | PubMed
- 47 Blanch AR, et al. (2006) Integrated analysis of established and novel microbial and chemical methods for microbial source tracking. *Appl Environ Microbiol* **72**:5915-5926 <https://doi.org/10.1128/aem.02453-05> | PubMed
- 48 Piracha Z, Saeed U, Khurshid A, Chaudhary WN (2014) Isolation and partial characterization of virulent phage specific against *Pseudomonas aeruginosa*. *Glob J Med Res* **14**:1-8
- 49 Chan BK, et al. (2016) Phage selection restores antibiotic sensitivity in MDR *Pseudomonas aeruginosa*. *Scientific Reports* **6** <https://doi.org/10.1038/srep26717> | PubMed
- 50 Gulino K, et al. (2020) Initial Mapping of the New York City Wastewater Virome. *mSystems* **5** <https://doi.org/10.1128/msystems.00876-19>
- 51 Olsen NS, Forero-Junco L, Kot W, Hansen LH (2020) Exploring the Remarkable Diversity of Culturable *Escherichia coli* Phages in the Danish Wastewater Environment. *Viruses* **12** <https://doi.org/10.3390/v12090986> | PubMed
- 52 Pruden A, Arabi M, Storteboom HN (2012) Correlation Between Upstream Human Activities and Riverine Antibiotic Resistance Genes. *Environmental Science & Technology* **46**:11541-11549 <https://doi.org/10.1021/es302657r> | PubMed
- 53 EasyMapMaker (no date) EasyMapMaker steps to Create a Custom Google map. <https://www.easymapmaker.com/>
- 54 Apprill A, McNally SP, Parsons RJ, Weber L (2015) Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat Microb Ecol* **75**:129-137 <https://doi.org/10.3354/ame01753>
- 55 Parada AE, Needham DM, Fuhrman JA (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* **18**:1403-1414 <https://doi.org/10.1111/1462-2920.13023> | PubMed
- 56 Schloss PD, et al. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**:7537-7541 <https://doi.org/10.1128/AEM.01541-09> | PubMed
- 57 Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl Environ Microbiol* **79**:5112-5120 <https://doi.org/10.1128/AEM.01043-13> | PubMed
- 58 Yilmaz P, et al. (2013) The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research* **42**:D643-D648 <https://doi.org/10.1093/nar/gkt1209> | PubMed

- 59 Quast C, et al. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**:D590-D596 <https://doi.org/10.1093/nar/gks1219> | PubMed
- 60 Ondov BD, Bergman NH, Phillippy AM (2011) Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12** <https://doi.org/10.1186/1471-2105-12-385> | PubMed
- 61 Community TG (2024) The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Research* **52**:W83-W94 <https://doi.org/10.1093/nar/gkae410> | PubMed
- 62 (2010) R: A Language and environment for statistical computing.
- 63 (2018) pheatmap: Pretty Heatmaps.
- 64 Arkin AP, et al. (2018) KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nature Biotechnology* **36**:566-569 <https://doi.org/10.1038/nbt.4163> | PubMed
- 65 (2010) FastQC: A quality control tool for high throughput sequencing data.
- 66 Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114-2120 <https://doi.org/10.1093/bioinformatics/btu170> | PubMed
- 67 Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**:824-834 <https://doi.org/10.1101/gr.213959.116> | PubMed
- 68 Swetnam TL, et al. (2024) CyVerse: Cyberinfrastructure for open science. *PLOS Computational Biology* **20**:e1011270 <https://doi.org/10.1371/journal.pcbi.1011270> | PubMed
- 69 Kieft K, Zhou Z, Anantharaman K (2020) VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8** <https://doi.org/10.1186/s40168-020-00867-0> | PubMed
- 70 Shang J, Peng C, Liao H, Tang X, Sun Y (2023) PhaBOX: a web server for identifying and characterizing phage contigs in metagenomic data. *Bioinformatics Advances* **3** <https://doi.org/10.1093/bioadv/vbad101> | PubMed
- 71 Nayfach S, et al. (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**:578-585 <https://doi.org/10.1038/s41587-020-00774-7> | PubMed
- 72 Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**:2068-2069 <https://doi.org/10.1093/bioinformatics/btu153> | PubMed
- 73 Bolduc B, et al. (2017) vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**:e3243 <https://doi.org/10.7717/peerj.3243> | PubMed
- 74 Shannon P, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**:2498-2504 <https://doi.org/10.1101/gr.1239303> | PubMed
- 75 Shang J, Jiang J, Sun Y (2021) Bacteriophage classification for assembled contigs using graph convolutional network. *Bioinformatics* **37**:i25-i33 <https://doi.org/10.1093/bioinformatics/btab293> | PubMed
- 76 Shang J, Sun Y (2022) CHERRY: a Computational methoD for accuratE pRediction of virus-pRokaryotic interactions using a graph encoder-decoder model. *Briefings in Bioinformatics* **23** <https://doi.org/10.1093/bib/bbac182> | PubMed
- 77 Nishimura Y, et al. (2017) ViPTree: the viral proteomic tree server. *Bioinformatics* **33**:2379-2380 <https://doi.org/10.1093/bioinformatics/btx157> | PubMed
- 78 Yu G, Smith DK, Zhu H, Guan Y, Lam TT (2017) -Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution* **8**:28-36 <https://doi.org/10.1111/2041-210X.12628>
- 79 Gregory AC, et al. (2022) MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. *Microbiome* **10** <https://doi.org/10.1186/s40168-022-01231-0> | PubMed

- 80 Bankevich A, et al. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**:455-477 <https://doi.org/10.1089/cmb.2012.0021> | PubMed
- 81 Brettin T, et al. (2015) RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* **5**:8365 <https://doi.org/10.1038/srep08365> | PubMed
- 82 Guo J, et al. (2021) VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9** <https://doi.org/10.1186/s40168-020-00990-y> | PubMed
- 83 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**:403-410 [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2) | PubMed
- 84 Bouras G, et al. (2023) Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics* **39** <https://doi.org/10.1093/bioinformatics/btac776> | PubMed
- 85 McNair K, Zhou C, Dinsdale EA, Souza B, Edwards RA (2019) PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* **35**:4537-4542 <https://doi.org/10.1093/bioinformatics/btz265> | PubMed
- 86 Chan Patricia P, Lin Brian Y, Mak Allisia J, Lowe Todd M (2021) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**:9077-9096 <https://doi.org/10.1093/nar/gkab688> | PubMed
- 87 Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* **32**:11-16 <https://doi.org/10.1093/nar/gkh152> | PubMed
- 88 Bland C, et al. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8** <https://doi.org/10.1186/1471-2105-8-209> | PubMed
- 89 Terzian P, et al. (2021) PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics* **3** <https://doi.org/10.1093/nargab/lqab067> | PubMed
- 90 Chen L, et al. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Research* **33**:D325-D328 <https://doi.org/10.1093/nar/gki008> | PubMed
- 91 Alcock BP, et al. (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Research* **48**:D517-D525 <https://doi.org/10.1093/nar/gkz935> | PubMed
- 92 Steinegger M, Söding J (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**:1026-1028 <https://doi.org/10.1038/nbt.3988> | PubMed
- 93 Larralde M, Camargo A (2023) Pyrodigal-gv: A Pyrodigal extension to predict genes in giant viruses and viruses with alternative genetic code. Github.
- 94 Larralde M (2022) Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software* **7** <https://doi.org/10.21105/joss.04296>
- 95 Larralde M, Zeller G (2023) PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* **39** <https://doi.org/10.1093/bioinformatics/btad214> | PubMed
- 96 Bouras G (2024) Dnaapler: A tool to reorient circular microbial genomes. *Journal of Open Source Software* **9**:5968 <https://doi.org/10.21105/joss.05968>
- 97 Cook R, et al. (2021) INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *Phage* **2**:214-223 <https://doi.org/10.1089/phage.2021.0007> | PubMed
- 98 Ondov BD, et al. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17** <https://doi.org/10.1186/s13059-016-0997-x> | PubMed

- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2025) Geomapping and PhiHD Sampling Metagenomes. NCBI BioProject. ID PRJNA1308632 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1308632>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2025) 16S Geomapping of Houston, TX. NCBI BioProject. ID PRJNA1309115 <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1309115>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) BSL02C1_1_clean. NCBI GenBank. ID PZ285931 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285931>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) BSL02C1_2_clean. NCBI GenBank. ID PZ285932 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285932>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) CSM01C1_clean. NCBI GenBank. ID PZ285933 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285933>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) MGB1X4C4_clean. NCBI GenBank. ID PZ285934 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285934>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) MYC16C1_clean. NCBI GenBank. ID PZ285935 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285935>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) MYC16C2_clean. NCBI GenBank. ID PZ285936 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285936>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) MYC30C1_1_clean. NCBI GenBank. ID PZ285937 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285937>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) MYC30C1_2_clean. NCBI GenBank. ID PZ285938 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285938>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS17X2C1_clean. NCBI GenBank. ID PZ285939 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285939>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS17X2C2_clean. NCBI GenBank. ID PZ285940 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285940>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS18C3_1_clean. NCBI GenBank. ID PZ285941 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285941>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS18C3_2_clean. NCBI GenBank. ID PZ285942 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285942>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS18C4_clean. NCBI GenBank. ID PZ285943 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285943>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS20C1_clean. NCBI GenBank. ID PZ285944 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285944>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS20C3_clean. NCBI GenBank. ID PZ285945 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285945>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS20C4_clean. NCBI GenBank. ID PZ285946 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285946>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS20C5_clean. NCBI GenBank. ID PZ285947 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285947>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26C1_1_clean. NCBI GenBank. ID PZ285948 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285948>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26C1_2_clean. NCBI GenBank. ID PZ285949 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285949>
- Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26C2_1_clean. NCBI GenBank. ID PZ285950 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285950>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26C2_2_clean. NCBI GenBank. ID PZ285951 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285951>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C1_1_clean. NCBI GenBank. ID PZ285952 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285952>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C2_1_clean. NCBI GenBank. ID PZ285953 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285953>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C2_2_clean. NCBI GenBank. ID PZ285954 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285954>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C2_3_clean. NCBI GenBank. ID PZ285955 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285955>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C2_4_clean. NCBI GenBank. ID PZ285956 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285956>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C2_6_clean. NCBI GenBank. ID PZ285957 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285957>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS29C1. NCBI GenBank. ID PZ285958 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285958>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS29C2. NCBI GenBank. ID PZ285959 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285959>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS37C1. NCBI GenBank. ID PZ285960 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285960>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS37C2_2_clean. NCBI GenBank. ID PZ285961 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285961>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) VP2. NCBI GenBank. ID PZ285962 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285962>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) VP3. NCBI GenBank. ID PZ285963 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285963>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) VP4. NCBI GenBank. ID PZ285964 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285964>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) VP1_GalvWB. NCBI GenBank. ID PZ285965 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285965>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) EUA3C1_clean. NCBI GenBank. ID PZ285966 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285966>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) MGB1X4C3_clean. NCBI GenBank. ID PZ285967 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285967>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) MGB1X4C1_clean. NCBI GenBank. ID PZ285968 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285968>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) 6X1C4_clean. NCBI GenBank. ID PZ285969 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285969>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C2_5_clean. NCBI GenBank. ID PZ285970 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285970>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) EUA02C1_clean. NCBI GenBank. ID PZ285971 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285971>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26X1C1_2_clean. NCBI GenBank. ID PZ285972 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285972>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS26C2_3_clean. NCBI GenBank. ID PZ285973 <https://www.ncbi.nlm.nih.gov/nuccore/PZ285973>

Do C, Salazar KC, Chang J, Clark JR, Terwilliger AL, Ruchhoeft P, Nicholls P, Maresso AW (2026) UCS37C2_1_clean. NCBI GenBank. ID PZ229304 <https://www.ncbi.nlm.nih.gov/nuccore/PZ229304>

Peer reviews

Reviewer #1 (Public review):

[Editors' note: this version has been assessed by the Reviewing Editor without an additional round of formal review from the original reviewers. The authors have addressed the comments raised in the previous round of review.]

Summary:

In the manuscript "Pathogen-Phage Geomapping to Overcome Resistance," Do et al. present an impressive demonstration of using geographical sampling and metagenomics to guide sample choice for enrichment in human-associated microbes and the pathogen of interest to increase the chances of success for isolating phages active against highly resistant bacterial strains. The authors document many notable successes (17!) with highly resistant bacterial isolates and share a thoughtfully structured phage discovery effort, potentially opening the door to similar geomapping efforts across the field. While the work is methodologically strong and valuable for the community, there are a few areas where additional clarification and analysis could better align the claims with the data presented.

Strengths:

- (1) The manuscript describes a well-executed and transparent example of overcoming a major obstacle in therapeutic virus identification, providing a practical success story that will resonate with researchers in microbiology and medicine.
- (2) Many phage researchers have anecdotally experienced a similar phenomenon, that a particular wastewater treatment plant always seems to have the pathogens you need. Quantifying this with metagenomics modernizes and adds evidence to this phenomenon in a way that could help researchers reproduce this success in a methodical way.
- (3) The methodology of combining environmental sampling, viral screening, and host-range analysis is clearly articulated and reproducible, offering a valuable blueprint for others in the field.
- (4) The data are presented with appropriate analytical rigor, and the results include robust sequencing and metagenomic profiling that deepen understanding of local viral communities.
- (5) The 17 successes yielding 35 phages have a lot of phylogenetic novelty beyond what the Tailor labs have typically found with previous methods.
- (6) The work highlights a practical and innovative solution to an increasingly important clinical problem, supporting the development of personalized antiviral strategies.

<https://doi.org/10.7554/eLife.109259.2.sa2>

Reviewer #2 (Public review):

Summary:

The manuscript by Do and colleagues aims to develop a workflow for isolating and identifying bacteriophages with potential applications in phage therapy against antibiotic-resistant pathogens. The workflow integrates geΦmapping as a strategy to identify potential

phage sources, Φ HD as a device for phage concentration, and R Φ as a phage library constructed from the initial sampling, resulting in the discovery of 36 new phages. The paper is overall interesting, and the proposed method appears robust and effective.

Strengths:

The methods proposed combined state-of-the-art strategies to solve an ever-increasing problem of antibiotic resistance. The methods are robust, and the controls are appropriate. The integration of environmental sampling, concentration strategies, and downstream genomic characterization is a clear strength and provides a potentially scalable framework for identifying candidate therapeutic phages. The manuscript is clearly written overall, and the results support the main conclusions.

Comments on revised version:

The manuscript has been adequately improved and adjusted according to the comments. There are minor points such as Table S10 is labelled in the top of the page as Table S11. Also, is a little unconventional to cite result figures and tables in the introduction.

For the question 10, regarding why some of the most abundant vOTUs in the 5L sample were not detected in the concentrate. The answer does not satisfy, as it focuses on why very low abundant vOTUs will not be detected, but the question is why some of the most abundant vOTUs were not detected. This does not affect the results or interpretation made.

<https://doi.org/10.7554/eLife.109259.2.sa1>

Author response:

The following is the authors' response to the original reviews.

Public Reviews:

Reviewer #1 (Public review):

(1) The central concept of geomapping as a broadly applicable strategy is wonderfully supported by the 17 successes documented in the paper. While this is actually, of course, a strength, the study does not include a comparative analysis across multiple sites with varying sampling outcomes for different bacterial types, which would be necessary to validate this claim more generally.

We thank the reviewer for the point, and it is well taken. We addressed this below, where we give a full discussion.

(2) Some elements, such as beta diversity comparisons and the metagenomics analysis of viral dark matter, would benefit from additional statistical analysis and clearer context.

The reviewer is quite correct as to the importance of bringing statistical analysis to our metagenomic analysis. To that end, we performed statistical analysis on our metagenomic datasets. We performed statistical analysis on our metagenomic datasets. We approached this using MetaPop to analyze viral metagenomic sequence data at the interpopulation (macrodiversity) level. MetaPop's macrodiversity analysis includes raw population abundance, normalized population abundance, and α -diversity calculations. With normalized population abundance tables, we were able to generate heatmaps to view feature-level distinction between samples and biomes. Furthermore, we were able to calculate β -diversity based on Bray-Curtis dissimilarity. PCoA was performed, and to assess robustness, 2,000 features were randomly subsampled and analysis repeated across 1,000 bootstrap iterations. Resulting ordinations were aligned to a reference with Procrustes alignment. Mean

coordinates and standard deviations were calculated for each sample, and scatter plots were generated. Supplementary Tables 6 and 8 and Supplementary Figure 4 have been added.

(3) Claims about therapeutic cocktails would be better framed as speculative and/or moved to the discussion section.

We thank the reviewer for their point, and it is well taken. Please see our more detailed response to this earlier in this reply.

(4) The manuscript could be strengthened by elaborating on the scope and composition of the phage and bacterial isolate collections, which are important for interpreting the broader significance of the findings.

We thank the reviewer for their point. We have added further details on the bacterial and phage isolate collections so the readers may draw the proper conclusions.

Reviewer #2 (Public review):

Weaknesses: >

While the authors acknowledge several limitations, some aspects require clearer framing or additional clarification. The proposed workflow focuses exclusively on aquatic environments as sources of phages, which may limit the diversity of hosts and phage types recoverable using this approach. Some interpretations, particularly regarding taxonomic classification and sampling saturation, would benefit from more cautious wording given current limitations in viral taxonomy and the observed data.

The reviewer makes an excellent point. To try and address this, we made several edits to the main text of the discussion section to reframe and add clarification to our limitations. We also mention the limitation of our strategy to aquatic environments. Lastly, we addressed the final sentence below.

Recommendations for the authors:

Reviewer #1 (Recommendations for the authors):

(1) To really demonstrate geomapping success would require more comparisons: choosing a variety of locations with and without high host levels and then analyzing the yes or no outcomes in terms of whether phages were found. This manuscript demonstrates 17 substantial and significant successes, very much worth sharing, but I am not sure it answers the central question posed in the title and abstract. This could potentially be accomplished through analysis of existing data related to the attempts, or it may be important to state this limitation in the discussion.

We thank the reviewer for their insightful comment on the generalizability of our geomapping strategy and their emphasis on adding more comparisons to support the claim. While we did not test the 17 bacterial isolates (XΦROs) across multiple sites with varied host levels, we did incorporate an initial, broad preliminary screening panel designed to assess the presence and diversity of phage against multiple pathogens and laboratory strains across different sampled environments after site comparisons with the geomap (Fig. 2G). In another instance, we did without the guidance of the geomap (Supp. Fig. 2I). In each case, highly polluted waters located in densely populated areas (wastewater, Brays Bayou, and Buffalo Bayou) had higher success rates in phage recovery compared to other less polluted sites (Clear Creek, Galveston seawater, Hamilton Pool Preserve, and Pedernales Falls). This trend is consistent with previous reports [16,51]. The screening step functioned as an initial comparison and allowed us to gauge phage availability across different genera of bacteria prior to a focused geomap-guided phage hunt. In the geomap-guided experiment, we compared a low host-availability control site (Clear Creek) with two high host-availability

sites (wastewater and Brays Bayou). More comparisons, especially incorporating even more sites with varied host availability, would strengthen the claim but not be invalid without it.

In an ideal situation, we would perform the exact additional experiments proposed by the reviewer. However, we ran into logistical issues. As it turns out, phage quantities in various sites vary with weather, specifically rainfall (a finding that we intend to touch on in a future manuscript). As such, this variable must be matched. Even with the long summers of Texas, we were unable to perform multiple, serial PhiHD runs at multiple sites with similar weather conditions. As such we rely on the strength of preliminary screening (16S and plating for phage) to provide a clear guide as to where to place PhiHD experiments.

We would still contend that the consistency between preliminary screening results and subsequent successful geomapping-guided discoveries (finding phages against 17 different isolates) gives sufficient evidence that geomapping is an effective strategy for identifying productive sampling sites. However, we acknowledge the excellent point of the reviewer and we have edited the main text to include this in the limitations, so that the reader may make their own evaluation.

(2) Line 44: 24% of infections >> perhaps better to describe as isolates, as many infections contain multiple isolates of different types. More background in terms of the number of infections in general, comprising the 24% on the bacterial side, and also a description of the 350 phages in terms of known hosts, would be very interesting to add.

This is an excellent suggestion that would add valuable background to the introduction and provide context for phage laboratories interested in the volume of cases handled by TAILOR and the isolates they receive. We changed the main text to include information regarding this suggestion. We also provided a three extra tables to show 1) TAILOR's number of distinct case counts with number of bacterial isolates received, 2) TAILOR's phage library, and 3) number of isolates with NO phage. Please see Supp. Tables 1-3.

(3) Line 92: To add a statistical test to the beta diversity comparisons beyond visual inspection of the location of the points on an ordination, I suggest adding a PERMANOVA.

The reviewer makes an excellent point, and we agree that visual inspection should be backed up by rigorous statistics where possible. As such, we performed an analysis of molecular variance (AMOVA, done in Mothur) to compare samples derived from brackish, sea, fresh, and sewage. AMOVA showed statistical significance between the samples and confirms our visual inspection of the β -analysis. The text has been altered to reflect AMOVA data.

(4) Line 112 (related to point 2 and Fig 1A): How many isolates in the collection for which 24% did not have a phage, and 35% were Pseudomonas?

Related to changes in point 2, we added a supplementary table to provide insight into the number of bacterial isolates without a phage. In total, 104 (24% of 435 isolates in the TAILOR library) have no phage, and 35 belong to *Pseudomonas aeruginosa*.

(5) Line 141: How is it known that Pseudomonas phage concentration increased by 95x? There could be unknowns / difficult to cultivate or phages without the right host. Consider describing as yields rather than the absolute concentrations.

We appreciate the reviewer's point that not all *Pseudomonas* phages are accounted for due to host specificity and potential unknowns. We agree completely that all we have are surrogates. We tracked the concentration of phages that infected our indicator strain *Pseudomonas aeruginosa* PAO1. We selected PAO1 because of its broad susceptibility profile and its role as a permissive host for isolating a wide range of *Pseudomonas* phages. While we acknowledge that not all *Pseudomonas* phages will be detected, PAO1 captures a wide breadth of

Pseudomonas phages, enabling consistent comparisons between samples. Our concentration changes reflect a within-sample comparison between unprocessed material and the processed material. Aligned with reviewer's concern, the values should not be taken as an estimate of absolute phage concentration in the samples. Rather, the values are method-dependent estimates of enrichment efficiency for PAO1-infecting phages. In tandem with PAO1-infecting phages, the concentration of other viral-like particles infecting other organisms is most likely increased with each concentration step. We have revised the entire manuscript to mention "phage yield," rather than associate an increase to a concentration.

(6) Line 224: 39.1% viruses> I believe this refers to vOTUs rather than viruses.

The reviewer is correct; we appreciate the catch! We corrected the text to "vOTUs."

(7) Lines 224-230: How does this relate to the expected ~70% Dark Matter?

As observed in many viral metagenomic studies, our dataset is also dominated by viral dark matter. Between 60.9% (outlier/singles from vCONTACT2) and 66.5% (unclassified from PhaGCN) of vOTUs are in this group. To proceed with caution, we edited the main text to draw attention to the large percentage of viral dark matter in our metagenomic dataset. Although a substantial fraction of vOTUs is unknown, the remaining identifiable sequences provide some biological context, enable validation of sampling strategies and comparative analyses between samples.

(8) Line 241: there are many perspectives on whether phage treatment should involve cocktails. If a phage is immunogenic and leads to antibody production that can neutralize other phages, one phage could ruin the game for others. Consider presenting this as a perspective, rather than a ground truth, and consider moving to discussion

This is a very insightful input on this perspective. Our intent was not to present this as a definitive conclusion, but rather to highlight a broader need for a more diverse phage library. This is not limited to phage cocktail generation. We changed the introductory sentence to this paragraph to encompass a broader need for phage diversity and succinctly lead into the next sentence.

(9) Figure 2b contains an R^2 value of 0.7, and 2c has $R^2=0.76$. Where does this come from? Maybe a PERMANOVA? Please describe in legend and/or methods+results.

Thank you for catching this! β -diversity calculations were based on Bray-Curtis dissimilarity. We have adjusted the methods and results to incorporate this information.

(10) The Rphi library is mentioned in several places, would be wonderful to have a bit more description of this collection.

We thank the reviewer for their sharp eye, we definitely wanted to ensure the reader understands the significance of this. We added some descriptor sentences to better highlight and introduce the R Φ -library.

(11) Consider adding a central success to the abstract, the fact that phages were found for 17 recalcitrant strains of various ESKAPE pathogens, yielding 35 phages after standard phage hunting and experimental evolution approaches had failed.

We appreciate the reviewer for their emphasis on highlighting the success of our manuscript and made the appropriate changes. We added altered the last sentence to the abstract, and added another sentence to summarize our success.

Reviewer #2 (Recommendations for the authors):

(1) Figures 1C and 1D require a more detailed description.

We thank the reviewer for noticing this. We have altered the figure legend to be more descriptive.

(2) Raw and assembled sequencing data should be submitted to a public repository, and the accession numbers should be provided.

We have uploaded the raw and assembled sequencing data to a public repository, and the accession numbers are provided in Supp. Table 9 and 12. For raw metagenomic shotgun sequences, BioProject accession is PRJNA1308632 (Supp. Table 2).

(3) Line 89: The text states that the rarefaction curves plateaued; however, by definition, a plateau implies that the curve no longer increases. In the presented data, all curves continue to rise at the final sampling point. This does not affect the conclusions but suggests that sampling saturation has not been fully reached.

This is a great observation by our reviewer. We agree with this point as the curves do not reach a complete plateau. We have revised the text to use more accurate language and clarify sampling depth.

(4) Figure 2D: The heatmap normalized by Z-score within the selected taxa may give a biased impression of enrichment of certain taxa in specific environments, when in fact it only indicates enrichment relative to the other pathogenic taxa included in the analysis.

The reviewer raises a great point, and we should have pointed this out directly. To avoid potential misinterpretations, we have revised main text to explicitly state that the heatmap displays relative enrichment to the other pathogenic taxa.

(5) Given the variable taxonomic resolution achieved by 16S rRNA sequencing (genus or family level), it would be important to highlight that some detected taxa include non-pathogenic members. For example, Vibrio is common in seawater, yet only a few species are pathogenic to humans.

We agree with this! We added a sentence to emphasize this point.

(6) Figure 2G: The color scale bar is uniform across all panels; please adjust for accurate comparison.

For more accurate comparisons between different samples and phage concentrations, we added a second color to assist with visualization.

(7) The PCoA figures should specify which distance metric was used.

We want to thank the reviewer for the catch, we should have mentioned that. Our PCoA was calculated based on Bray-Curtis dissimilarity. We have adjusted the main text and methods section to mention it.

(8) Figure 3: The meaning of the colors in panels A and B should be clarified.

We thank the reviewer for their keen eye. We changed the figure and figure legend to clarify. The colors on the map and PCoA represent influents from various wastewater treatment plants around Texas.

(9) The manuscript jumps from Supplementary Figure 2 to Figure 6. In general, the order and referencing of supplementary materials are confusing. Supplementary tables and figures should not be intercalated within the same file.

We thank the reviewer for their patience and apologize for the confusion. This occurred as we had multiple revisions to the manuscript and we did not update the sequence of the

figures. To address the reviewer's comment, we separated the supplementary tables and figures into apart. We also ensured that each main and supplementary figures and table were mentioned sequentially in the main text.

(10) It is unclear why some vOTUs were observed in the 5 L collection but not in the concentrated sample (10/24; Supplementary Figure 3E). One would expect that the most abundant vOTUs in the 5 L sample should also be easily detected in the concentrate.

The reviewer brings up a fantastic point. One would certainly expect that the most abundant vOTUs in the 5L samples would also be detected in the concentrated sample.

We have several suspicions as to why several vOTUS were not detected in our concentrated samples. Because we used shallow shotgun metagenomic sequencing, as compared to deep sequencing, we may have obscured our ability to detect and quantify low-abundance taxa. Consequently, dominant taxa occupying a large portion of sequencing reads may have masked the detection of rarer species/vOTUs. Lower sequencing depth results in fewer total reads per sample and reduced sensitivity for rare, infrequent species/vOTUs to be detected. When their abundance falls below detecting limits, they may appear absent from a data set.

Furthermore, we reached out to Novogene, who we outsourced for library preparation and shotgun metagenomic sequencing. According to Novogene, not all genetic material in a sample is used during their library preparation. The maximum amount of DNA to build a PCR-free metagenomic library at each time is approximately 1.5 µg of DNA. Although we submitted 184.4 µg of DNA (from the 400L-concentrate) and 25.6 µg of DNA (from the 5L-sample), we suspect only a fraction of the material was used for library preparation and subsequently sequenced. This may have limited the representation of low abundance vOTUs.

(11) Line 227: The statement that "but only 33.5% of viruses could be classified to the family-level" requires caution. Since the traditional Siphoviridae, Podoviridae, and Myoviridae families were abolished, many viruses currently lack family-level classification. Therefore, this taxonomic level may not be ideal for assessing novelty, as many viruses closely related to known types remain unassigned.

We appreciate the reviewer for bringing up this topic. We recognize the current limitation in the viral metagenomic landscape. Many viruses lack-family level classification due to ICTV taxonomic restructuring and the lack of reference genomes present in a database. A large fraction of viral sequences constitute "viral dark matter." From a single metagenomic dataset, viral dark matter ranges from 60-90% of vOTUs. Within our own dataset, it is also dominated by viral dark matter. Between 60.9% (outlier/singles from vCONTACT2) and 66.5% (unclassified from PhaGCN) of vOTUs are in this group. Although a substantial fraction of vOTUs is unknown, the remaining identifiable sequences provide some biological context, enable validation of sampling strategies and comparative analyses between samples. To complement vCONTACT2 results, we utilized PhaGCN to classify each vOTU as a means to compare taxa derived from each sampled biome from one another and not to assess novelty of the metagenomic dataset. We aimed to provide measurable and interpretable context to our metagenomes. However, due to the substantial variability and uncertainty in the field and our dataset, we revised the text to highlight the large fraction of unclassified sequences and their implications.

(12) Supplementary Figure 5: The legend does not clearly explain the two inner rings. One may correspond to GC skew, but this should be explicitly stated.

Well spotted! We have made the appropriate corrections.

(13) Line 325: The reference to "50 mL samples" is unclear-please specify which samples this refers to.

<https://doi.org/10.7554/eLife.109259.2.sa0>