Development and application of bioinformatic tools for analyzing phage replication cycles and host interactions in health and disease

Dissertation

Dissertation der Fakultät für Biologie

der Ludwig-Maximilians-Universität München

Xue Peng

Munich, Nov. 2024

Diese Dissertation wurde angefertigt
unter der Leitung von PD Dr. Jürgen Lassak
im Bereich von Microbiology, Department Biology
an der Ludwig-Maximilians-Universität München

Ersgutachter/in:	PD Dr. Jürgen Lassak
Zweitgutachter/in:	Prof.Dr. Li Deng
Tag der Abgabe:	13.11.2024
Tag der mündlichen Prüfung:	07.03.2025

EIDESSTATTLICHE RKLÄRUNG

Ich versichere hier mit an Eides statt, dass meine Dissertation selbständig und ohne unerlaubte Hilfsmittel angefertigt worden ist.

Die vorliegende Dissertation wurde weder ganz, noch teilweise bei einer anderen Prüfungs kommission vorgelegt.

Ich habe noch zu keinem früheren Zeitpunkt versucht, eine Dissertation einzureichen oder an einer Doktorprüfung teilzunehmen.

München, den 13.11.2024

XUE PENG

Table of content

Abstract (English):

The phage-host interaction (PHI) is a dynamic and complex process influenced by multiple factors, such as extensive genetic and morphological diversity among bacteriophages. Prophages, as the direct interaction between phages and their bacterial hosts, can replicate alongside bacteria and facilitate their adaptation to the environment by providing additional functions. This study investigated the phage replication cycles—lytic, lysogenic, and chronic—to uncover their roles in phage therapy and patients with gut disorders. To enhance accuracy in identifying phage replication cycles, particularly the chronic cycle, I developed RepliDec, a computational tool capable of predicting lytic, lysogenic, and chronic lifecycles. RepliDec addresses existing limitations of available computational tools that often misclassify chronic phages within complete phage genomes and the fragmented sequences obtained from sequencing technology. I first assessed RepliDec's performance on complete phage genomes, and it outperformed all other tools, achieving the highest scores across all four metrics: sensitivity (86.76%), accuracy (85.57%), F1 score (87.31%), and Matthews correlation coefficient (MCC) (74.00%). Additionally, RepliDec demonstrated the best performance in all metrics for simulated phage contigs, achieving an accuracy of 77.04% and an F1 score of 70.69%. Furthermore, I developed an integrated pipeline, RepliDec+, designed for use in complex microbial communities, such as the human gut. Subsequently, by employing RepliDec+, I re-evaluated the prevalence of temperate phages in patients diagnosed with inflammatory bowel disease (IBD). I discovered that temperate phages are significantly more prevalent among IBD patients experiencing severe symptoms compared to healthy controls. Furthermore, I investigated the presence of temperate phages in four commercially available phage cocktails. It is crucial to note that temperate phages should be excluded from these cocktails due to their propensity for harboring virulent genes, which may be transmitted to bacteria, potentially leading to adverse outcomes. My analysis indicated the detection of several temperate sequences containing integrases within the PYO cocktails. Moreover, a minimal level of bacterial contamination was noted in this study. The insights gained through RepliDec and Replidec+ advance our understanding of prophages and contribute to microbiome studies as well as the development of bacteriophage-based therapeutic strategies.

Identifying phage-host pairs is essential in PHI studies, with experimental methods, such as plaque assays, remaining the gold standard for this recognition. Investigating PHI in the gut

environment is challenging due to its complex network, which involves mammalian hosts, bacteria, viruses, and fungi. To address this complexity, I utilized a modified version of viral tagging (VT) and whole genomic sequencing to establish potential phage-bacteria associations from cross-infection samples. I identified 607 viral clusters (VCs) and 208 bacterial taxa, contributing to the determination of the phage-bacteria association network across three disease conditions: ulcerative colitis (UC), early-stage colorectal cancer (CRCE), and advanced-stage colorectal cancer (CRCA). Thousands of phage-bacteria associations were detected within each disease condition. This approach enhances our understanding of gut microbiome dynamics and phage-host interactions in both health and disease.

This work advances our understanding of phage-host interactions, particularly in the context of the human gut microbiome. The findings demonstrated the key role of temperate phages with IBD and phage therapy. Furthermore, the use of the modified viral tagging (VT) method to establish phage-bacteria associations paves the way for future studies to explore microbial interaction pairs on a larger scale. Investigating the bacteria and their associated viral communities may provide new insights into the application of phage therapy as a viable treatment option.

List of figures

List of tables

List of abbreviations

The following list includes frequently used abbreviations. All other abbreviations are explained in the main text.

AMR	Antimicrobial Resistance
CD	Crohn's Disease
CRC	Colorectal Cancer
CRCA	Advance Stage of Colorectal Cancer
CRCE	Early Stage of Colorectal Cancer
EPTC	Eliava Phage Therapy Center
IBD	Inflammatory Bowel Disease
PC	Protein Cluster
PHI	Phage-Host Interaction
UC	Ulcerative Colitis
VC	Viral Cluster
VT	Viral Tagging

1. Introduction

1.1 Bacteriophages

1.1.1 Bacteriophage history and current status

Viruses are the most abundant organisms in the world. The total abundance of viruses was estimated as 10^31 on the planet (Comeau et al., 2008; “Microbiology by Numbers,” 2011), much more than the bacteria in the human body (Shkoporov & Hill, 2019). Most of these viruses are bacteriophages or phages (Breitbart et al., 2007; Camarillo-Guerrero et al., 2021).

One hundred years ago, William Twort first discovered bacteriophages while propagating smallpox vaccinia virus (Twort, 1961; Keen, 2015). He found some “glassy and transparent” spots in bacteria contamination plaques. However, no clear answer can explain this observation. Several years later, Felix d’Herelle had similar observations and realized it was a new virus type. This new virus could kill bacteria and form a transparent dead bacteria zone (d’Herelle, 1961). They were named as bacteriophage (phage) (d’Herelle, 1961). Since then, research on bacteriophages has raised the curtain.

Initially, research on bacteriophages attracted little attention, and only a few articles were published each year (Figure 1-1). Until 2000, there was a steep increase in published articles related to bacteriophages. In 2023, the number of published articles was three times that during 1968-1999. This increase indicates that phages have begun to draw people’s attention, and their importance in ecology and the environment has been realized and studied widely.

Even though bacteriophages were discovered a hundred years ago, their vast diversity, unique infection cycles, and highly diverse genomes continue to pose significant challenges to researchers aiming to unveil their secrets. Thanks to new technology, such as sequencing technology, the cultural dependence method is bypassed, and the research on bacteriophages is accelerated. However, many “dark matters” in bacteriophages (Clooney et al., 2019; Santiago-Rodriguez & Hollister, 2022) still need to be investigated in the future.

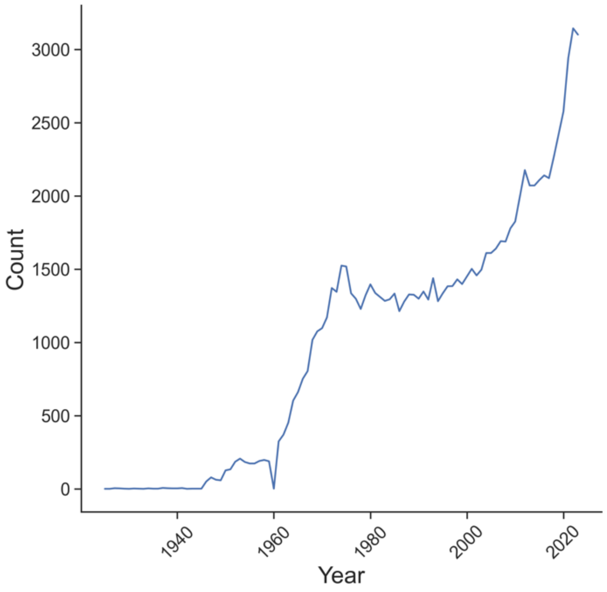


Figure 1-1 **The number of published reseach articles related to phages.** Searching for "Phage" in PubMed (https://pubmed.ncbi.nlm.nih.gov/).

1.1.2 The significance of studying bacteriophages

Bacteriophages are ubiquitous organisms worldwide, and their impact on the whole living system is neglectable. For example, bacteriophages identified from ocean or lake environments can significantly impact productivity and nutrient cycling among water-living organisms (Hurwitz & U'Ren, 2016). They possess genes related to sulfur (Hesketh-Best et al., 2023; Kieft, Zhou, et al., 2021), carbon (Hesketh-Best et al., 2023; Sanmukh et al., 2015), and nitrogen (Waldbauer et al., 2019) which indicate they have the potential to take part in these metabolic cycles. Moreover, these metabolic cycles might take part in global climate change (Focardi et al., 2020). Researchers suggest that marine viruses should be considered part of ocean climate models to increase the accuracy of climate change prediction (Danovaro et al., 2011; McDonald, 2016).

Viruses/bacteriophages can also manipulate human organic sulfur metabolism (Kieft, Breister, et al., 2021). More and more phage studies have demonstrated that they play an essential role in directly or indirectly influencing human health, such as Inflammatory Bowel Disease (IBD) (Norman et al., 2015), *Clostridium difficile* infection (CDI) (Zuo et al., 2017), diabetes (Ma et al., 2018; Zhao et al., 2017), hypertension (Han et al., 2018), asthma (Megremis et al., 2023) and

Parkinson's disease (PD) (Tetz et al., 2018). They also impact the nervous system (Jędrusiak et al., 2023) and immune system (Champagne-Jorgensen et al., 2023).

1.1.3 Bacteriophage biology

Bacteriophages are viruses that can infect bacteria or archaea. They exhibit significant diversity in both their morphology (Figure 1-2 A) and genomes. According to the latest ICTV Taxonomy Release (VMR_MSL38_v3; https://ictv.global/vmr), bacteriophages are mainly from 10 classes: *Tokiviricetes*, *Caudoviricetes*, *Faserviricetes*, *Malgrandaviricetes*, *Huolimaviricetes*, *Vidaverviricetes*, *Leviviricetes*, *Ainoaviricetes*, *Tectiliviricetes*, *Laserviricetes*. This new classification system is genome-based virus taxonomy (Koonin et al., 2020). Virions belonging to the *Leviviricetes* class and *Microviridae* family both feature a unique spherical morphology with icosahedral symmetry. However, there are distinct genetic and structural variations between them. Phage belonging to *Leviviricetes* are all positive single-strand RNA (+ssRNA) viruses, encoding three +ssRNA virus core proteins: a maturation protein (MP), a coat protein (CP), and a catalytic subunit of an RNA-directed RNA polymerase (RdRP) (Callanan et al., 2021). In contrast, *Microviridae* phages are characterized by genomes consisting of circular positive sense ssDNA molecules, typically around 4.4–6.1 kb in length (Roux et al., 2012). Moreover, phages belonging to the *Corticoviridae* family are also a family of icosahedral viruses and their genomes consist of highly supercoiled circular double-stranded DNA, typically around 10 kb in length (Oksanen & ICTV Report Consortium, 2017). Their virions possess internal lipid membranes, which are absent in *Leviviricetes* and *Microviridae*.

The high diversity of bacteriophages poses a challenge to category them easily. The new bacteriophage taxonomy system offers finer resolution than older systems (Koonin et al., 2020), allowing for more genomic differentiation among the phage family and genus. However, given the high diversity observed among phages, relying solely on taxonomy information may not suffice for their precise characterization. Their diversity manifests in various aspects such as genomic composition, genome size, virion particle size, morphology, replication cycle, and host range.

Based on their genomic composition, phages can be categorized into four groups: (a) single-stranded RNA phage, (b) double-stranded RNA phage, (c) single-stranded DNA phage, or (d) double-stranded DNA phage. Their genomes also can be linear, circular, or segmented, and genome size ranges from the ~3,300 nucleotide ssRNA viruses of Escherichia coli to 735 kb circularized huge phage genome (Al-Shayeb et al., 2020; Hatfull & Hendrix, 2011).

Moreover, there is an essential group of single-stranded DNA (ssDNA) bacteriophages whose virion is long, thin, and filamentous in shape. They belong to the *Inoviridae* family and have a major coat protein helically organized around supercoiled, circular ssDNA. The length of Inoviridae bacteriophage virions depends on their genome size, ranging from 600 to 2500 nm with a diameter of 6–10 nm (Knezevic et al., 2021). Recent studies have verified that *Inoviridae* is prevalent in prokaryotic bacteria and archaea and identified 10,295 inovirus-like sequences from microbial genomes and metagenomes, which largely expand the current inovirus database (Roux, Krupovic, et al., 2019). In addition, phylogenetic analysis reveals the high diversity within the *Inoviridae* family, leading to suggestions for reclassifying *Inoviridae* into a viral order and establishing six new families (Roux, Krupovic, et al., 2019).

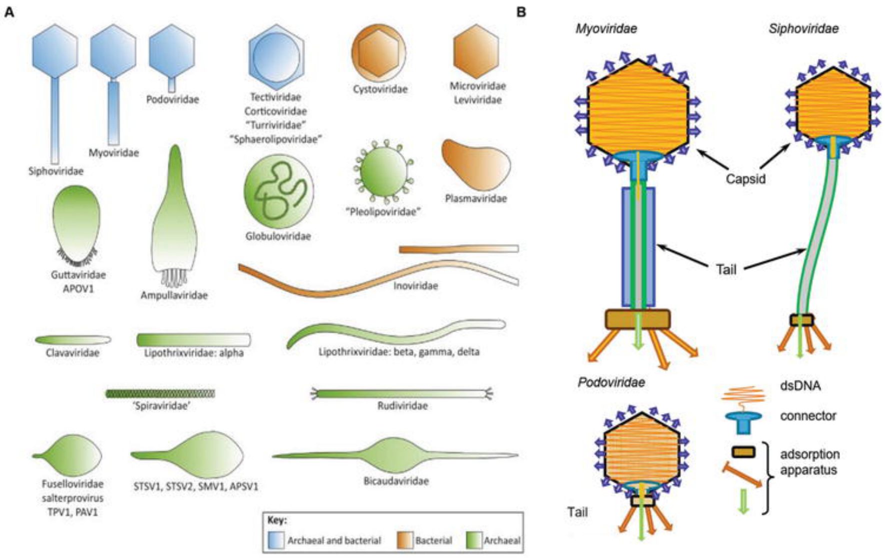


Figure 1-2 **Bacteriophages morphology.**
(A) Representation of prokaryote bacteriophage morphotypes. (B) Members of the Caudoviricetes class. Figure from chapter *Bacteriophages: Their Structural Organization and Function*. by White Helen E. and Orlova Elena V in book *Bacteriophages - Perspectives and Future (E. White & V. Orlova, 2020)*.

A study examining 5568 phages in the electron microscope found that about 96% of them are tailed phages (Ackermann, 2007). Moreover, gut virome studies also verified that tailed phages are the most abundant in gut virome (Camarillo-Guerrero et al., 2021; Z. Cao et al., 2022; Ma et al., 2018). Tailed bacteriophages are an essential part of studying phage biology. Tailed phages are all double-stranded DNA (dsDNA) phages belonging to the order *Caudovirales* (abolished according to ICTV 2023). All the tailed phages were described in class *Caudoviricetes* in the new system, and abolition of the morphology-based families *Myoviridae*, *Podoviridae*, and *Siphoviridae* (Turner et al., 2023). However, phages in this class mainly have three different morphologies according to the type of tails. (a) *Podoviridae* virions have a short, non-contractile tail and their size is about 20×8 nm; (b) Virions belonging to *Siphoviridae* have a long, non-contractile, thin, and flexible tail. The diameter of their capsids is approximately 60 nm, while their tails measure around 150×8 nm. (c) *Myoviridae* virions have a long and complex contractile tail, which is relatively thick (~16nm) compared to *Siphoviridae*. Their heads are prolate icosahedra (elongated pentagonal bipyramidal antiprisms), and their size is about 111×78 nm.

In addition to differences in genomic composition and virion particles, the diversity of bacteriophages is evident in their infection processes, including the replication cycle and host specificity. One notable characteristic is their host specificity. It is a prevalent but insufficiently tested belief that bacteriophages possess a narrow host range, capable of infecting various strains within the same species (Fong et al., 2021). For example, *Pseudomonas phage BrSP1* can infect 19 *Pseudomonas aeruginosa* strains (De Melo et al., 2019). This distinctive characteristic enables bacteriophages to function as an effective and safe alternative to antibiotics for treating multidrug-resistant bacteria. Additionally, with the increasing discovery and isolation of various bacteriophages, researchers have identified specific phages with a relatively broad host range, allowing them to infect numerous species or even bacteria from different genera (De Melo et al., 2019; Göller et al., 2021; Islam et al., 2023; Tétart et al., 1996). In one study, researchers investigated host ranges of staphylococcal phages isolated from wastewater. A diverse host panel of 123 bacteria from 32 different species was used, including 29 *Staphylococcus* species (117 strains), two *Macrococcus* species (4 strains), and one *Enterococcus* species (2 strains). The study revealed that ten out of ninety-four unique phages exhibited a uniform host range and were isolated from multiple species. Furthermore, only sixty-four unique phages were isolated on a single occasion (one plaque), while thirty-two were isolated between two and fifteen times (Göller et al., 2021). Another critical example is bacteriophage T4, which can infect a range of *Escherichia coli* strains and some closely related bacteria such as *Shigella* (Tétart et al., 1996). From this evidence, we can infer that phage host specificity is not that strict.

The relatively extensive host spectrum is advantageous for phage therapy, as broad-host-range phages can infect and eradicate a more diverse array of pathogen strains or related species. Nonetheless, this extensive host spectrum presents a challenge in investigating the phage-host relationship. Despite advancements in the comprehension of phage-bacteria interactions, predicting a phage's host range solely based on genomic characteristics is difficult due to the considerable genetic diversity observed among phages. Moreover, phages infecting the same host may exhibit significant genomic diversity. For example, a genome comparison analysis between 27 phages infecting *Staphylococcus aureus* reveals extensive genome mosaicism in their gene maps. The 27 S. aureus phages can be organized into three groups based on genome size, gene map organization, and comparative nucleotide and protein sequence analysis (Kwan et al., 2005). These studies demonstrate that phage host specificity is complex and influenced by a variety of factors without clear, predictable patterns in their genomic features. The broad host range, combined with the highly mosaic genome, makes phage host prediction more complex and poses a challenge to phage classification by merely using an infected host. Further research is needed to identify potential patterns and improve our understanding of phage host range determinants.

1.1.4 Phage infection cycle

Bacteriophages can infect bacteria or archaea hosts. In general, bacteriophages need several steps to accomplish the infection process: (a) attachment, (b) penetration, (c) biosynthesis, (d) maturation, and (e) release (Sausset et al., 2020). The infection starts with the bacteriophage attaching to the surface of host cells via specific receptors. Bacteriophage receptor-binding proteins (RBPs) display high specificity for these receptors (Costa et al., 2022). This unique attachment mechanism limits the phage host range. After attachment, the phage injects its genome through the host cell wall and membrane while the rest of its components, such as the capsid and tail, remain outside the host. Once inside, phages adopt different strategies to interact with the bacterial genome. Some phages will degrade the bacterial chromosome and hijack the host machinery to synthesize early enzymes, coat proteins, and replicate the bacteriophage genome. These components are then used to assemble new virions. After maturation, phage proteins such as holin or lysozyme lyse the host cell, releasing the new virions. However, some phages will integrate their genome into the bacterial chromosome and replicate with the host together. Until some conditions (e.g., ultraviolet light exposure or chemical exposure) stimulate the host cell, the integrated phage genome (prophage) will excise from the bacteria genome. After that, they will utilize the host machinery, produce the new virions, and then kill the host. Different replication strategies are vital factors when characterizing phages.

1.1.5 Current application of phage in clinical

Before the discovery of penicillin, the first antibiotic, phages were used as medicine for ten years (Chanishvili, 2012). Phage therapy is an essential application of bacteriophages with a long history, but its global application is not yet widespread. The historical evolution of phage therapy shows that political issues, personal rivalries, and ongoing disputes have greatly hindered its advancement (Summers, 2012). The first documented clinical application of phage therapy occurred in 1919 at the Hôpital des Enfants-Malades in Paris, where it was successfully utilized to treat four pediatric cases of bacterial dysentery (Chanishvili, 2012; d'Herelle, 1961; Summers, 2012). The first field trials used bacteriophages as a prophylactic against avian typhosis (*Salmonella gallinarum*) in rural France in 1919 ("LIMITATIONS OF BACTERIOPHAGE THERAPY," 1931). Until 1930, it has been used worldwide in humans, but scientists have different opinions about phage therapy. Some think it's a promising treatment, while others think it is over-sold (Summers, 2012). However, the broad spectrum and constant dosage of antibiotics, coupled with the mixed success of phage therapy in clinical trials, have resulted in a predominant preference for antibiotics within Western medical practice (Chanishvili, 2012; Lin et al., 2017). As a result, phage therapy experienced a significant decline in acceptance and utilization in Western medicine after the introduction of pharmaceutical antibiotics.

Along with the widespread use of antibiotics, the overprescribing and overusing of antibiotics lead to severe problems, such as increased antimicrobial resistance (AMR), particularly the emergence of antibiotic-resistant bacteria strains (Llor & Bjerrum, 2014). These bacteria will cause more severe symptoms, prolonging the illness, increasing mortality rates, and raising the risk of complications and hospitalization (Llor & Bjerrum, 2014). Moreover, a report from the World Health Organization (WHO) shows that there has been extensive overuse of antibiotics worldwide during the COVID-19 pandemic, which may have worsened the "silent" spread of AMR (World Health Organization, 2024). According to a study, bacterial AMR was directly responsible for 1.27

million global deaths in 2019 and contributed to 4.95 million deaths (Murray et al., 2022). AMR is becoming one of the serious public health problems and development threats worldwide. Several actions have been taken globally to address this issue, such as the One Health Approach and the Global Action Plan on Antimicrobial Resistance (World Health Organization, 2023). Additionally, efforts are being made at the academic level to promote research and the development of alternative therapies like bacteriophage therapy and antimicrobial peptides (Huan et al., 2020). A successful clinical use of phage therapy to treat severe multi-antibiotic-resistant *Acinetobacter baumannii* infection in the U.S. at the University of California San Diego reignited interest in phage therapy (Hitchcock et al., 2023; LaVergne et al., 2018).

Kutter et al. (Kutter et al., 2010) have described the previous clinical phage therapy trials in detail. More and more studies have been conducted (Fowoyo, 2024; Swenson et al., 2024), while the regulation and acceptance of phage therapy vary among countries. Poland and Soviet republics, notably the Georgian and Russian, have been pivotal in advancing the field of phage therapy. The Eliava Phage Therapy Center (EPTC) in Tbilisi, Georgia, and the Hirszfeld Institute of Immunology and Experimental Therapy in Poland are the most famous institutions in phage therapy. EPTC treated 400 patients with phages in 2018 and 20 to 30 patients with cystic fibrosis per month, according to the staff report at a phage conference in 2019 (Swenson et al., 2024). The Hirszfeld Institute reported treating 1307 patients with phage preparations from 1987 to 1999 (Żaczek et al., 2020). The Israeli Phage Therapy Center (IPTC) has treated 20 patients primarily for bone infections since its establishment from 2018 to 2023 (Onallah et al., 2023). In the United States, it has received approval for application within veterinary and agricultural contexts, targeting animals and plants. However, the Food and Drug Administration (FDA) has not approved bacteriophage-based human clinical use products (Swenson et al., 2024). However, the FDA has been accepting more innovative clinical trial designs (Swenson et al., 2024; US Food and Drug Administration, 2019). According to the records listed on clinicaltrials.gov (access date: 2024-09-25), 45 phage therapy clinical trials are listed in total. Seven of them were completed, and 25 were undergoing. Since 2020, 37 clinical trials have been initiated to explore the efficacy and safety of phage therapy. Despite the relatively limited application of phage therapy in clinical trials, there has been a noticeable increase over the last five years compared to prior periods. This trend highlights the growing interest in and potential of phage therapy as a viable treatment option within the medical community.

Phage therapy has been used in several human clinical trials to treat skin and soft tissue infections (Gupta et al., 2019; Jault et al., 2019), respiratory tract infections (Aslam et al., 2019; Lebeaux et al., 2021), urinary tract infections (Kuipers et al., 2019; Zaldastanishvili et al., 2021), and Gastrointestinal infections (Federici et al., 2022). Many bacteriophage clinical trials currently registered on clinicaltrials.gov are focused on applying bacteriophages to treat skin and soft tissue infections, accounting for 21.7%, and gastrointestinal infections, representing 16.2% (Hitchcock et al., 2023). These clinical trials mainly focus on treating infections caused by ESKAPE pathogens. ESKAPE stands for *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and species of *Enterobacter*. Among these pathogens, *Staphylococcus aureus*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* are the most common in bacteriophage clinical trial records. *Acinetobacter baumannii*, *Enterococcus faecium*, and species of *Enterobacter* are less frequently encountered in the records (Hitchcock et al., 2023). These ESKAPE pathogens are prone to mutating into antibiotic-resistant strains and are hard to be killed by antibiotics. For chronic nonhealing burn wound infections, Phage cocktail PP1131 has been used and PP1131 decreased the bacterial burden in burn wounds at a slower pace than the standard of care (Jault et al., 2019).

Moreover, gastrointestinal infections mainly result from bacteria, with increases attributed to the rise of antibiotic-resistant strains. Without effective treatments, phage therapy has been proposed as a clinical alternative to help restore intestinal balance. Researchers developed a five-phage therapy that successfully suppresses *Klebsiella pneumonia*, which is associated with worsening inflammatory bowel disease (IBD) (Federici et al., 2022). And tested it in an artificial human gut and with healthy volunteers, showing safety, viability, and potential for treating non-communicable diseases by targeting antibiotic-resistant gut bacteria. Also, Intralytix's EcoActive™ bacteriophage therapy targeting adherent-invasive *Escherichia coli* (AIEC) in Crohn's disease patients has entered a Phase 1/2a clinical trial (registered under clinicaltrials.gov NCT03808103) and recruited volunteers with inactive Crohn's disease at the Icahn School of Medicine at the Mount Sinai Hospital in New York, NY, and the Johns Hopkins University in the Baltimore, MD metro area.

1.1.6 Application of phage in argiculture and food industry

Despite its long history of use in medicine, phages have also been utilized in various non-clinical applications (Elfadadny et al., 2024; García-Cruz et al., 2023). In the food industry, phages can be used as preservatives and antibacterials in packaged products to prevent bacterial contamination (Ranveer et al., 2024). In the agriculture industry, they can be used as a safe and effective approach to pest control, especially for those who rely on endosymbionts for essential physiological functions, such as aphids.

Phage also plays a crucial role in the aquaculture industry. Bacterial infections, for instance, greatly contribute to fish morbidity and mortality, frequently resulting in disease outbreaks and significant economic damage. At present, multiple bacterial strains have been reported to infect fish, including *Aeromonas septicemia*, *Edwardsiellosis*, *Columnaris*, *Streptococcosis*, and *Vibriosis* (Irshath et al., 2023). While antibiotics and chemotherapeutics can be used to treat diseases caused by these bacteria, they have several notable drawbacks, including drug resistance and safety concerns for consumers and the environment. Bacteriophages can be a safe alternative to address these issues, and oral administration of a phage cocktail is the most suitable method of application in fish (Soliman et al., 2019).

In our daily lives, phages can also be helpful as disinfectants for decontaminating or cleaning surfaces, such as in hospitals. Phages are natural predators of bacteria, and their host specificity makes them safe and effective at killing bacteria, especially those within biofilms that are resistant to chemical disinfectants. So, phage disinfectants will be more environmentally friendly and safe than chemical disinfectants (Elfadadny et al., 2024).

In addition, phages can be utilized to protect the environment, such as combating global warming and treating wastewater (García-Cruz et al., 2023). For example, methane ranks as the second most prevalent greenhouse gas after CO2, with a notable portion originating from methanogenic archaea found in the gut microbiome of cattle bred for human consumption. Phages targeting eubacteria and methanogenic archaea could also serve as an effective strategy by reducing the numbers of these eubacteria and methanogenic archaea, leading to a decrease in methane emissions (García-Cruz et al., 2023). A similar strategy can also be applied to CO_2.

1.1.7 Human virome and metagenomics

Advancements in sequencing technology enable the discovery and characterization of bacteriophages at the population level without the need for culture. Following the extraction of DNA from samples, a sequencing library must be prepared based on the sequencing technologies. Subsequently, the fragmented DNA sequences library are sent for sequencing. Short-read sequencing typically generates reads ranging from 50 to 200 base pairs (bp). Thousands of short reads can be analyzed using two strategies: reference-based and reference-free methods. The reference-based approach requires a reference sequence; it involves mapping the short reads to the reference to ascertain the relative abundance of the sequences. When a reference or an appropriate reference database is unavailable, a de-novo assembly strategy must be applied, utilizing mathematical algorithms to connect the short reads into longer contigs or scaffolds. Depending on the specific investigative objectives, various downstream analyses can be conducted. In the case of bacteriophages, this often involves identifying prophages within bacterial genomes, distinguishing viral contigs from all assembled contigs, predicting the host range and replication cycle of viral contigs, and characterizing the taxonomy of these unknown contigs. Sometimes, it is necessary to predict and annotate the functions of genes in order to investigate their roles within metabolic pathways.

Although several tools have been developed for each analysis, they have already been discussed in other articles (Khan Mirzaei et al., 2021; Knight et al., 2018). All these tools can be classified into three types: alignment-based, alignment-free, or hybrid. Alignment-based tools utilize information within the public databases and align it with query sequences; the one exhibiting the highest score or similarity will be selected as a result. However, there exists a significant amount of "dark matter” in the virome (Santiago-Rodriguez & Hollister, 2022), which complicates the analysis. Additionally, this approach lacks the ability to handle novel phages. Alignment-free tools utilize machine learning or deep learning methods to transform biological signals into a high-dimensional space and forecast outcomes using mathematical calculations. The downside of this approach is its lack of interpretability, and it relies heavily on a large training dataset. Some tools integrate both strategies, but they do not eliminate the drawbacks of either approach.

1.1.8 Viral genomics in virome and metagenomics

The most common topic for investigating phages in virome and metagenomic data is to identify viral contigs from the assembled contigs, including identifying the viral regions or prophage regions within bacterial contigs and differential viral contigs from contamination contigs. There are many tools available for distinguishing viral from microbial sequences, typically they can classified into three groups based on their methods: homology-based, homology-independent, and hybird. Homology-based approaches utilize alignment similarities with known databases as criteria to ensure the viral source, such as MetaPhinder (Jurtz et al., 2016) and VirSorter (Roux et al., 2015). Nevertheless, the high diversity of phages, combined with the plasmid sequences found in bacterial genomes and auxiliary metabolic genes carried by phages, makes it difficult to clearly distinguish between phages and bacteria. Furthermore, the absence of comprehensive phage databases hinders the accurate prediction of prophage regions within microbial sequences. The homology-independent approach utilizes gene content and genomic structural features or the frequencies of DNA "words" (i.e., k-mers), transforming them into high-dimensional space, and subsequently integrates machine learning or deep learning techniques, such as DeepVirFinder (Ren et al., 2020), PPR-Meta (Z. Fang et al., 2019) and PhiSpy (Akhter et al., 2012). While machine learning methods that rely on nucleotide composition and gene content would not be impacted much by viral database bias, they tend to be confused by any unusual sequence as plasmids or eukaryotic genome fragments (Guo et al., 2021). Some tools, VIBRANT (Kieft et al., 2020) and VirSorter2 (Guo et al., 2021), combine both approaches to aim for improved predictions.

Once the viral contigs are identified, the subsequent analyses vary based on the specific research purpose. Typically, viral contigs longer than 10 kb are ideally chosen for future analysis (Camarillo-Guerrero et al., 2021). This will result in information loss from contigs shorter than 10 kb. To handle this, there are two ways; first is to construct the viral Operational Taxonomic Unit (vOTU) by clustering these viral contigs based on genomic similarity. However, vOTU does not represent a true species or strains; rather, it is a collection of similar viral contigs, which can potentially lead to inaccuracies in diversity assessment. Second is binning these contigs into metagenome-assembled genomes (MAGs) based on their co-abundance or genomic content information. Viral contigs belonging to the same species or strain will be grouped into bins, with each bin treated as a distinct “real” phage. However, mosaic genomic features of phages may pose challenges in binning; therefore, it is essential to conduct a post-evaluation for each bin. Currently, METABAT (D. D. Kang et al., 2019), PHAMB (Johansen et al., 2022) , and vRhyme (Kieft et al., 2022) can be used to binning viral contigs.

Once the viral contigs or the viral MAGs are obtained, we can predict the open reading frame (ORF) from them using Prodigal (Hyatt et al., 2010) and Glimmer (Salzberg et al., 1998). Then, the proteins can be mapped to a public function database to find the possible function annotation. There are multiple databases that can be used, such as PFAM (Mistry et al., 2021), KEGG (Kanehisa & Goto, 2000), NCBI blast NR database (O’Leary et al., 2016), CDD (M. Yang et al., 2020), UniProt (The UniProt Consortium et al., 2023), COG (Galperin et al., 2021), PHROG (Terzian et al., 2021), CARD (Alcock et al., 2023), ARDB (Liu & Pop, 2009), VOGDB (Trgovec-Greif et al., 2024) and pVOGs (Grazziotin et al., 2017). Despite lots of databases available, a large proportion of hypothetical proteins have unknown functions. It is essential to explore new strategies to accelerate the process of characterizing these unknown functions.

1.2 Phage replication cycle and identification

1.2.1 Replication cycle

There are four types of replication cycles, including lytic, lysogenic, pseudolysogeny and chronic (Olszak et al., 2017; Sieiro et al., 2020) (Figure 1-3). (a) In the lytic cycle, virulent phages hijack the host's cellular machinery to make new copies of the phages, and the phage progeny will be released through the lysis of the host cell. (b) The lysogenic cycle happens when a phage enters into a stable symbiosis with its bacterial host by integrating its genome into the host chromosome or is maintained as a plasmid and replicates with the bacterial host (Weinbauer, 2004). The host and phage in such a relationship are called a lysogen and a prophage, respectively. Prophages can enter the lytic cycle spontaneously at a low frequency and in response to external stressors such as those that trigger DNA damage or the SOS response (Clokie et al., 2011; Hobbs & Abedon, 2016; Mirzaei & Maurice, 2017). (c) Chronic cycle, common among filamentous phages, e.g. phage M13, is a special replication cycle whereby virions are produced and continuously released without killing the host. Additionally, there is one case known as pseudolysogeny (Łoś

& Węgrzyn, 2012; Olszak et al., 2017; Sieiro et al., 2020), which is associated with the starvation stress of the bacterial host and can occur in both virulent and temperate phages. In this state, the phage genome remains inactive after injecting its genome into host cells, and the viral genome has no degradation. The phages do not produce new virions nor replicate in synchronization with the host cell. However, once nutrients become sufficient, they will begin to enter either the lytic or lysogenic cycle.

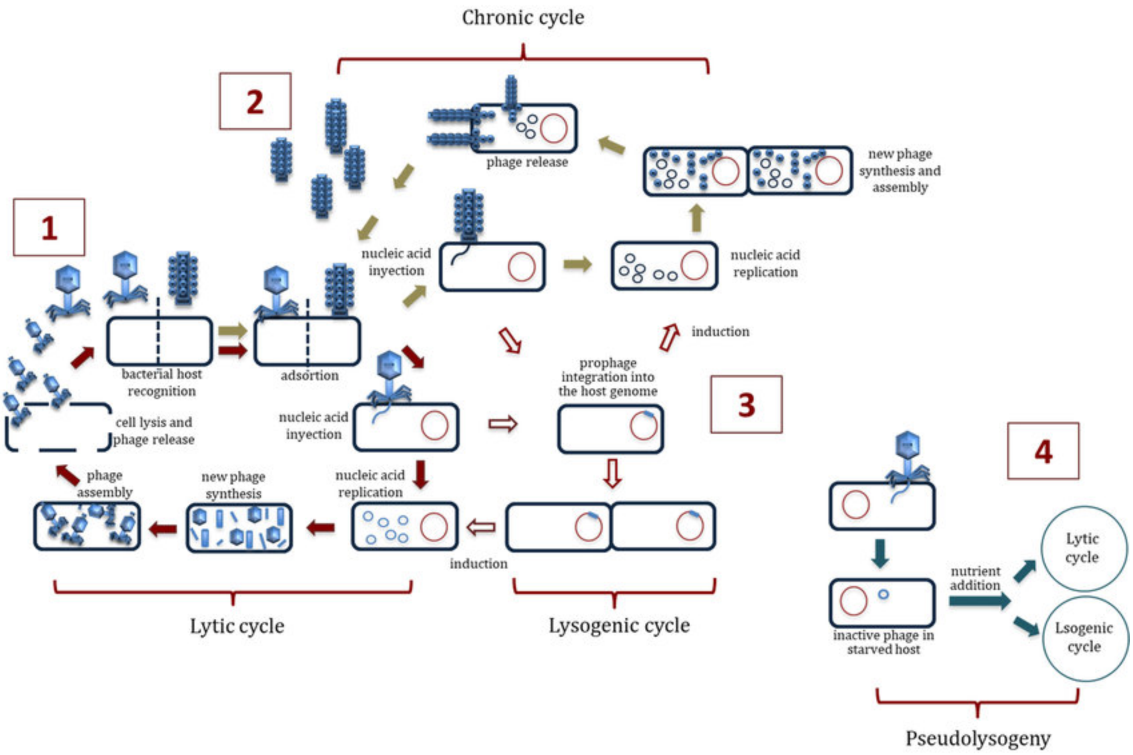


Figure 1-3 **Phage replication cycles.**
This figure originates from Sieiro et al. (Sieiro et al., 2020).

1.2.2 Identification of phage replication cycle

There are some ways to identify the phage replication cycle. If the phage can be cultured in the lab, the replication cycle can be determined by plaque assay. In plaque assay, virulent phages might form clear plaques, while temperate phages might form turbid plaques (Jurczak-Kurek et al., 2016). It's essential to verify the observed plaque morphologies through phage genomics to ensure that lysogeny-related genes are absent from the phage genome before concluding the replication cycle. Experimental methods do not always work in certain situations. For example, phages cannot be cultured and isolated in the lab, and sometimes phages are identified using sequencing technologies. Shotgun sequencing bypasses the need for culturing, but it only provides genomic-level information and generates hundreds or even thousands of viral contigs in complex ecosystems. Determining the replication cycle of these viral contigs through experimental methods is often time-consuming and ineffective, especially since most of them are unculturable. These limitations compel us to seek alternative solutions to these challenges.

A practical way to tackle these challenges is by using computational techniques to predict the viral replication cycle. By accurately forecasting this cycle, we can gain crucial insights into virus-host interactions within intricate ecosystems and emphasize the virome's significance in human health and disease (Clooney et al., 2019; Khan Mirzaei et al., 2020; Mirzaei & Maurice, 2017; Roux, Adriaenssens, et al., 2019). Currently, various tools can predict phage replication using phage genomes (Table 1-1). However, utilizing computational software to predict the replication cycles of uncultivated phages has some limitations (Table 1-1). Generally, these tools use one or both prediction strategies: alignment-based and alignment-free. Alignment-based strategy relies on a set of marker genes or proteins that commonly exist in temperate phages. For example, PhageLead (Yukgehnaish et al., 2022) uses five different temperate marker genes, including Integrease, ParA, Cro/CI, immunity repressor, and antirepressor. PHACTS (McNair et al., 2012) randomly selected 600 query proteins as markers, and BACPHLIP (Hockenberry & Wilke, 2021) uses 206 lysogeny marker proteins. The alignment-based strategy can achieve high accuracy, but not all phages contain these markers due to their significant variability and high mosaic organization in phage genomes (Hatfull, 2008; Hatfull & Hendrix, 2011). Also, the fragmented phage assemblies from sequencing data pose a challenge in detecting these markers. Alignment-based strategy might confidently predict the replication cycle well on the long and complete genomes (Hockenberry & Wilke, 2021). However, the alignment-free strategy uses k-mer or machine learning algorithms to transform the DNA data into high-dimension space, such as PhageAI (Tynecki et al., 2020), PhagePred (Song, 2020), DeePhage (Wu et al., 2021), and PhaBOX/PhaTYP (Shang et al., 2023). The alignment-free strategy considers the information of the entire input sequence, not limited to the completeness or the marker genes of the phage genome. PhageAI relies on Word2Vec with the Ship-gram model and linear Support Vector Machine to predict the phage's replication cycle (Tynecki et al., 2020). PhagePred uses k-mer frequencies as the sequence feature to evaluate the d_2^S dissimilarity measures between novel viral sequences and two types of replication cycles (lytic/lysogenic) in phages (Song, 2020). DeePhage uses a convolutional neural network to detect signatures associated with lytic/lysogenic cycles in viral sequences and predict their replication cycle (Wu et al., 2021). PhaBOX/PhaTYP uses Bidirectional Encoder Representations from Transformer (BERT) to learn the protein composition and associations and a contextualized embedding model from natural language processing (NLP) for replication cycle classification (Shang et al., 2023). An alignment-free strategy can address the limitations of alignment-based methods. However, its effectiveness depends on having access to large and diverse datasets for training the model. If the training datasets are too small or lack diversity, it can lead to overfitting during the training process, undermining the model's accuracy and robustness. However, the training dataset used by most of these tools is mainly from 2 sources: (a) Mavrich’s dataset which contains about 1551 RefSeq phages (507 empirically established; the rest are predicted from 206 “temperate phage” domain) (Mavrich & Hatfull, 2017) or (b) RefSeq genomes predicted using alignment-based strategy established by Mavrich (Mavrich & Hatfull, 2017). According to Mavrich’s research, the computational prediction strategy exhibits a specific error rate whereby the predicted lifecycle data conflicts with the empirical lifecycle data in 4% of the empirical dataset (Mavrich & Hatfull, 2017).

In current tools, inoviridae is classified as temperate/virulent instead of chronic. This misclassification may impact the accuracy of chronic phage categorization in downstream analysis. Moreover, without a standardized benchmark dataset, these tools rely on different testing datasets, making comparing performance with previous results difficult.

Table 1-1 Software comparison of phage replication prediction tools

SOFTWARE	AIM	PREDICTION	PLATFORM	ALGROTHOM	TRANING DATASET	TEST DATASET	DISADVANTAGE
PHACTS (2012)	complete phage genomes; partial genome: at least a third of a phage's proteome is needed to accurately predict the lifecycle of a phage.	Temperate | Virulent	stand-alone (python)	protein homology alignment + supervised Random Forest classifier	protein datasets	148 temperate phages and 79 virulent phages from PHANTOME database	Dataset out of date; Large dataset will take very long time
PHAGEAI (2020)	complete phage genomes;	Temperate | Virulent | Chronic	web; stand-alone (python)	Machine Learning (SVM) + Natural Language Processing (Word2Vec with Ship-gram model)	278 virulent and 174 temperate phages (most from RefSeq) using bioinformatic prediction via lysogenic factors	54 virulent and 30 temperate phages using bioinformatic prediction via lysogenic factors	Have query or search limitation (100 query/day), but can query for more; not suitable for large dataset such as virome dataset

PHAGEPRED (2020)	phages which derived from the metagenomic data	Temperate|Virulent	stand-alone (python)	k-mer frequencies between teperate and virulent	Refseq sequence predicted in Mavrich and Hatfull (2017): 617 Virulent; 330 Temperate	Refseq sequence predicted in Mavrich and Hatfull (2017): 253 Virulent; 72 Temperate	novel virues phage might hard to preidcted; k-mer size k and orders of Markov models can markedly impact the performance, due to the high variability and highly mosaic organization of phages, need to choose longer length of k-mer and higher order of Markov chain; small training dataset
BACPHLIP (2021)	complete phage genomes; fragmented or partially assembled genomes is likely to be substantially degraded	Temperate | Virulent	stand-alone (python)	conserved protein domains + Random Forest classifier	60% Dataset (634): 1,057 phages with empirically established lifecycles that were collected by Mavrich & Hatfull (2017); may contain errors or inconsistencies in reporting	40% Dataset (423)	BACPHLIP relies on finding protein domains within the input genome. If a genome is only 50% complete, the lack of lysogeny-associated proteins may be due to the fact that the phage genome is virulent or it may simply be because the relevant domains are encoded within the missing genome segments.

DEEPHAGE (2021)	viral fragments from metagenome and metavirome	Temperate|Virulent	stand-alone (Linux operating system)	convolutional neural network + “one-hot” encoding form	80% Dataset-1: 77 virulent phages and 148 temperate manually curated with credible lifecycle annotations from the dataset of McNair et al. (dataset from PHACTS) ; 100% Dataset-2: 1,211 virulent and 429 temperate phage genomes from Refseq predicted using a bioinformatic method from Mavrich & Hatfull (2017)	20% Dataset-1; real virome data: bodily fluids in the bovine rumen using “excision," “integration," or “lysogeny" viral genes	predicted each contig, can not use for binning; need MATLAB dependencies; Linux only
PHABOX/ PHATYP (2023)	complete phage genomes; viral fragments from metagenome and metavirome	Temperate|Virulent	stand-alone (Linux operating system)	natural language processing (NLP) + Bidirectional Encoder Representations from Transformer (BERT)	3474 Refseq phage genomes	Dataset-1: 77 virulent phages and 148 temperate phages from PHANTOME database (PHACT test dataset) Dataset-2: 1211 virulent phages and 429 temperate phages (DeePhage Dataset2) Dataset-3: 2291 phage contigs assembled from metagenomic data from Liang G et.al.	predicted each contig, can not use for binning; Linux only

REPLIDEC	complete phage genomes; viral fragments from metagenome and metavirome	Temperate | Virulent | Chronic	stand-alone (python)	Naive Bayes classifier + alignment based (integrase + excisionase)	1,159,179 proteins including 419,983 proteins were predicted from 4,126 viral genomes and 739,196 prophage proteins from 21,134 lysogenic prophages identified in 14,922 bacterial and archaeal genomes.	Dataset-1: 470 RefSeq viral genomes from Mavrich which their replication cycle was experimented validated and consistent with the bioinformatic prediction results; Dataset-2: 610 represented viral genomes (can present 2920 genomes) which show at most 95% similarity between each other also not similar to RefSeq and even some prophage sequences. lifecycle predicted using bioinformatic method from Mavrich and homology to prokaryotic sequences;	needs multiple threading when handle very large dataset (n>3000 sequences)

1.3 Phage cocktail

1.3.1 Phage selection and cocktail design

While phage therapy shows great promise, various obstacles hinder its adoption. Strict regulations and approval processes vary across countries, and limited clinical data—particularly the absence of large, double-blind, placebo-controlled trials—complicate the assessment of phage treatment's effectiveness. Other issues also need to be addressed.

The primary and most critical factor is to the production and standardization of phages. Phages utilized in phage cocktails must adhere to several criteria. Firstly, phages should exhibit lytic replication cycle and must not contain virulent factors that release toxins. Furthermore, they should lack the capacity to transduce antibiotic resistance genes or virulent factors among bacteria hosts. Additionally, phages should possess a relatively broad host range. The mixtures of phages should target different host receptors to ensure the efficacy of the phage cocktail. Moreover, they should be cultivable for industrial applications. Except that, they should exhibiting a slow *in vivo* virion decay rate so that they have a long-lasting effect (Bull & Gill, 2014).

Thanks to sequencing technology, bacteriophage discovery and characterization can be accelerated. However, this is insufficient for phages used in phage therapy. Currently, there are 12,197 phage sequences listed in the NCBI Viral Genome Resource (Brister et al., 2015) (access date: 20240928). However, this number is far smaller than the estimated 10^31 phages in the natural environment (Comeau et al., 2008; “Microbiology by Numbers,” 2011).

Besides, most of the phages listed in NCBI were discovered by the SEA-Phages program at the Howard Hughes Medical Institute (Hitchcock et al., 2023). Although different phage discovery protocols can yield varying phages, they generally involve four main steps: isolation, purification, amplification, and characterization (Figure 1-4). Initially, phage isolation commences with the filtration of environmental samples such as wastewater, hospital waste, fecal samples, etc. And bacteria resources were removed from the environmental samples. Subsequently, the resultant filtrate is amalgamated with a targeted bacterial culture in an enrichment process. This step ensures the efficacy of phages in augmentation against the designated host bacteria. After enrichment, a spot assay is used to detect the presence of phages on a bacterial lawn. Phages extracted from the resulting phage plaque are mixed with a neutral buffer, and serial dilutions are conducted prior to additional plating to ensure clear phage plaques with a consistent appearance. The phages obtained from the plaque need to be characterized using sequence technology. After getting the raw reads from the sequencing machine, assemble them into the draft or complete genomes. Compare to the phage genomes in the public database to ascertain the novelty of the phage. Conclusively, these phages are cataloged in bacteriophage banks and international phage directories, making them available for subsequent applications.

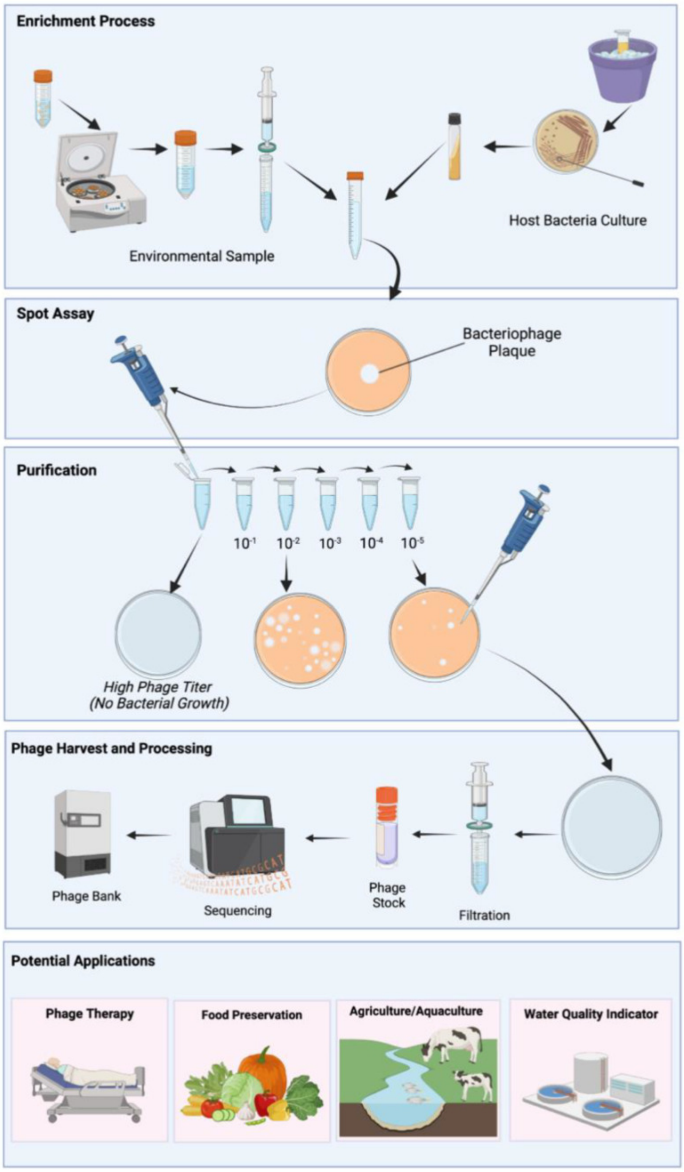


Figure 1-4 **Process of bacteriophage isolation for therapeutic use.**
Figure from the article “Current Clinical Landscape and Global Potential of Bacteriophage Therapy” (Hitchcock et al., 2023).

Once new bacteriophages have been discovered and isolated, it is essential to enforce stringent quality control protocols prior to their application in clinical settings. These protocols are vital to reducing contamination risks, identifying mutations in phage strains, and evaluating the characteristics of the phages to confirm their suitability for therapeutic purposes. If a phage is not strictly lytic, genetic engineering might be required to adjust it for safe use in clinical applications (Park et al., 2017). This matters because temperate phages can pass on virulence factors and genes linked to antibiotic resistance to bacterial hosts, which may worsen a patient's condition. The variability among phage strains, risks of contamination, and the necessity for phages to stay stable during production and storage present further challenges for large-scale manufacturing.

Moreover, phage therapy is a treatment that is highly specific to each patient and challenging to manufacture, unlike antibiotics. Phages specifically target specific bacterial hosts, making it challenging to achieve high efficiency without proper titers and identifying the bacterial strain responsible for the infection. While combining multiple phages into a cocktail can address phage specificity.

Moreover, the human microbiome differs considerably from person to person, with variations also present across different diseases (Amos et al., 2021; D.-Y. Kang et al., 2023; Mills et al., 2022). The relationship between phages and the microbiome remains unclear, as the microbiome may interact with phages, reducing their effectiveness. Individual immune responses can differ even when patients receive the same treatment. For more effective outcomes, a cocktail with a mix of broad host spectrum bacteria, or personalized phage cocktail, or a combination of phage therapy with antibiotics may be considered. However, personalized phage therapy is not well-suited for large-scale production and might not yield significant profit compared to commercially manufactured antibiotics. Consequently, pharmaceutical companies have limited their investments in this field. Since phages are natural entities, controlling the market for phage treatments poses a challenge for any company, which could also decrease the motivation for pharmaceutical companies to invest in research and development. The use of phage treatment in clinical settings is more stringent than in the food and agricultural industries.

There are several commercial phage cocktails available that can be used in clinical treatment. (

Table 1-2). And most of them are treatment the antibiotic resistance strains. Moreover, most of them are from Eliava BioPreparations in Georgia (https://phage.ge/en/products) and Microgen in Russia (https://www.microgen.ru/en/products/bakteriofagi). The investigation into phages is increasingly thorough Europe. PhageEU, founded by Proteon Pharmaceuticals, JAFRAL, and the PTC Phage Technology Center, was registered on July 11, 2024, and is seated in Brussels. PhageEU aims to influence the political and regulatory landscape in Europe, ensuring that the full benefits of phages are harnessed for EU citizens, patients, and farmers. The "DZIF Translational Phage Network" (DZIF TransPhage-Net) has undertaken the responsibility of uniting researchers, medical professionals, and veterinarians who share an interest in phages within Germany while improving interactions with pharmaceutical producers and regulatory agencies. ESCMID Study Group for Non-Traditional Antibacterial Therapy (ESGNTA) brings together professionals focused on advancing bacteriophages and non-traditional therapies for infectious diseases.

Table 1-2 List of commercial phage cocktails

Name	treatment	Manufacturer
Phago-Pyo bacteriophage	diseases caused by *Staphylococcus* spp., *Streptococcus* spp., different types of *Escherichia coli*, *Pseudomonas aeruginosa*, *Proteus mirabilis* and *Proteus vulgaris*.	Eliava BioPreparations, Tbilisi, Georgia
Phago-Intesti bacteriophage	diseases caused by Shigella (*Shigella flexneri* serotype 1,2, 3, 4 and *Shigella sonnei*), Salmonella (*S. paratyphi* A, *S. paratyphi* B, *S. typhimurium*, *S. enteritidis*, *S. choleraesuis*, *S. oranienburg*), different types of *Escherichia coli*, *Pseudomonas mirabilis, Pseudomonas vulgaris*, Staphylococcus (*S. aureus*), *Pseudomonas aeruginosa* and Enterococcus spp.	Eliava BioPreparations, Tbilisi, Georgia
Phago-Enko bacteriophage	diseases caused by *Salmonella typhimurium*, *Salmonella enteritidis*, *Salmonella heidelberg*, *Salmonella newport*, *Salmonella cholerae*, *Salmonella oranienburg*, *Sal- monella dublin* and *Salmonella anatum; Shigella flexneri* (serovars 1, 2, 3, 4) and *Shigella sonnei* (six different sero- vars), enteropathogenic *E. coli, S. aureus, Staphylococcus epidermidis, and Staphylococcus saprophyticus.*	Eliava BioPreparations, Tbilisi, Georgia
Phago-SES bacteriophage	diseases caused by Staphylococci (*S. aureus, S. epidermidis* and *S. saprophyticus*); Streptococci (*Strepto- coccus pyogenes, Streptococcus sanguis, Streptococcus sali- varius* and *Streptococcus agalactiae*) and different serotypes of enteropathogenic *E. coli* serovars.	Eliava BioPreparations, Tbilisi, Georgia

Phago-Staph bacteriophage	diseases caused by *Staphylococcus* spp.	Eliava BioPreparations, Tbilisi, Georgia
Phago-FERSISI bacteriophage	diseases caused by *Staphylococci and Streptococci*	Eliava BioPreparations, Tbilisi, Georgia
EcoActive™ bacteriophage	*adherent-invasive Escherichia coli (AIEC) in Crohn's disease patients*	Intralytix, Columbia, MD, USA
VRELysin™ bacteriophage	vancomycin resistant *Enterococci colonization* in the gastrointestinal tract	Intralytix, Columbia, MD, USA
WPP-201	Venous leg ulcers caused by *Pseudomonas aeruginosa*, *Staphylococci aureus*, and *Escherichia coli*	Intralytix, Columbia, MD, USA
LBP-EC01	treat uncomplicated urinary tract infections and other infections caused by antibiotic-resistant *Escherichia coli*	Locus Biosciences, Morrisville, NC, USA
E.coli-Proteus bacteriophage	diseases caused by bacteria *Proteus* and enterotoxigenic *Escherichia coli*	Microgen, Moscow, Russia
Pyobacteriophage polyvalent purified	diseases caused by bacteria *Staphylococcus, Streptococcus, Proteus, Pseudomonas aeruginosa, Klebsiella pneumoniae, Escherichia coli*	Microgen, Moscow, Russia
Streptophage	diseases caused by *Streptococcus*	Microgen, Moscow, Russia
Piophage	diseases caused by *Staphylococci*, *Streptococci*, *Enterococci*, *Proteus*, *Klebsiella pneumoniae* and	Microgen, Moscow, Russia

	Klebsiella oxytoca, *Pseudomonas aeruginosa* and *Escherichia coli*	
Sextaphage	diseases caused by *staphylococci, streptococci, Proteus, Klebsiella, Pseudomonas aeruginosa and Escherichia coli*	Microgen, Moscow, Russia
Intestiphage	diseases of the gastrointestinal tract caused by bacteria of *dysentery, Salmonella, Escherichia coli, Proteus, enterococci, staphylococci, Pseudomonas aeruginoza*	Microgen, Moscow, Russia

1.4 Phage-bacteria interaction and viral tagging

1.4.1 Phage-bacteria interaction

The interaction between phages and bacteria is a complex and dynamic network that influences microbial evolution, ecology, and biotechnology applications. The interactions within ecosystems can be explained through two ecological concepts: "kill-the-winner" (KTW) and "piggyback-the-winner" (PTW) (X. Chen et al., 2021). In the KTW model, bacteriophages target a community's most abundant and successful bacterial populations. This concept is based on the idea that rapidly growing and dominant bacterial populations become prime targets for phages. As a result, bacteriophages kill dominant bacteria, controlling the population and promoting biodiversity. In the PTW model, bacteriophages enter a lysogenic replication cycle with their bacterial hosts instead of killing the most successful bacterial populations. The phage integrates its DNA into the bacterial genome, becoming a prophage, and remains dormant within the bacterial cell. This allows the virus to "piggyback" on the success of the dominant bacterial population without destroying it (Silveira & Rohwer, 2016). These two hypotheses separately describe the community-level interaction between bacteria and phages with different lifecycles.

However, when narrowing it down to the genetic level, the phage-host interaction is more complex than these two hypotheses. When phages are attached to the bacteria's cell surface, the phage recognizes the bacteria via specific receptors, which are unique to different bacterial species or strains. This means the phage only infects particular bacteria strains. After injection, there is a battle between phages and bacteria. The bacteria will try to eliminate the phage DNA, while the phages attempt to lyse the bacteria genome or integrate into the host genome. So, there are many anti-phage systems, including single-gene systems like AbiH, Lit, NixI, and BstA, as well as multi-gene systems such as CRISPR-Cas, BREX (bacteriophage exclusion), DISARM (defense islands system associated with RM), Dnd, and Ssp systems (Georjon & Bernheim, 2023). Out of 21,364 fully sequenced bacterial genomes, 78% encode more than two defense systems, with significant variations between strains (Tesson et al., 2022). This indicates that the antiphase system is highly diverse and prevalent among bacteria.

Phage-host interactions (PHI) are not always antagonistic interactions. They can also be mutualism and parasitism. Mutualism is for temperate phages or prophages; they integrate their genome into the bacteria's genome and replicate. Phages also contain some virulent factors and auxiliary metabolic genes (AMG), which can provide the bacterium with beneficial traits, such as increased virulence, antibiotic resistance, or metabolic capabilities (Hurwitz & U'Ren, 2016; Kieft, Zhou, et al., 2021; Thompson et al., 2011; Waldbauer et al., 2019). The lysogenic conversion of prophages is sometimes associated with horizontal gene transfer (HGT). This process helps transfer genes among bacterial community members, playing a critical role in bacterial populations' genetic diversification and adaptability by acquiring new capabilities. Besides, the prophages can also enhance bacterial superinfection immunity, protecting the bacteria from other potentially harmful viral attacks (Abedon, 2022). This mutual defense benefits the host while allowing the prophage to persist in a stable environment. Furthermore, the interaction between chronic phages, such as filamentous phages, and their hosts is characterized as parasitic. This is because chronic phages can produce new virions without resulting in the death of the bacteria cell (Liang & Radosevich, 2020).

The phage-host interaction is a complex, multi-layer interaction network involving many small molecules and proteins working together. However, many unsolved "dark matters" exist in the relationship, and the mechanism underlying the essential interactions is still unknown. More work needs to be done in the future.

1.4.2 Approaches to predict phage–host interactions

Some experimental approaches can be used to identify the phage-host interaction (PHI), such as spot assays and plaque assays, liquid assays, viral tagging, microfluidic PCR, PhageFISH, single-cell sequencing, and Hi-C sequencing (Edwards et al., 2016). Culturing phages can be experimentally challenging. This is because phages may require specific conditions to grow. Additionally, most bacteria have not yet been cultured, limiting hosts' availability. Furthermore, some phages, such as those that establish a lysogenic infection cycle, may be challenging to observe and detect. Currently, some computational tools for predicting phage host relationships exist, such as iPHoP (Roux et al., 2022), PhiSpy (Akhter et al., 2012), PHISDetector (Zhou et al., n.d.), VIBRANT (Kieft et al., 2020), etc. The predictions typically derive from molecular features of coevolution or an arms race between bacteriophages and their bacterial hosts. This includes exact matches to reference viral or host genomes, matches to host-encoded CRISPR spacers, and sequence composition analyses like oligonucleotide profiles (Edwards et al., 2016; Versoza & Pfeifer, 2022). For some community sequencing data, abundance profiles of phages and bacteria can be used to predict phage-host relationships according to their predator-prey relationship (Edwards et al., 2016; Versoza & Pfeifer, 2022).

1.4.3 Viral tagging in phage-host interaction

Viral tagging is a direct and targeted technique (Džunková et al., 2019; Marbouty et al., 2021). It uses a fluorescent dye to stain virus-like particles, which mix with the bacteria and co-incubate for some time. Then, flow cytometry is used to sort bacterial cells with a fluorescently labeled phage attached. Subsequently, DNA is extracted from the sorted viral-tagged cells for whole genome sequencing or single-cell sequencing (Deng et al., 2014; Džunková et al., 2019). The viral tagging technique can reveal novel host phage pairs, including identifying a number of novel pairings with *Synechococcus* in marine and 363 unique host–phage pairings in human stool (Deng et al., 2014; Džunková et al., 2019). This approach enables us to decipher relationships between phages and their hosts in complex natural environments. However, using original viral tagging to obtain the whole picture of PHIs from the complex environment could be time-consuming and involve hard labor work. Advancements in sequencing technology, particularly integrating single-cell sequencing with viral tagging, enable a comprehensive understanding of PHIs.

1.5 Overview of gastrointestinal disorders

Digestive diseases are disorders of the gastrointestinal (GI) tract, and 60 to 70 million Americans are affected by digestive diseases (National Institutes of Health & US Department of Health, 2009). According to the European Federation of Crohn's and Ulcerative Colitis Associations (EFCCA), there are 10 million people worldwide diagnosed with inflammatory bowel diseases (IBDs)(European Federation of Crohn’s and Ulcerative Colitis Associations (EFCCA), n.d), and among them, an estimated 2.39 million IBD patients in America(Lewis et al., 2023). IBD can be classified into two types, ulcerative colitis (UC) and Crohn’s disease (CD), based on different symptoms, inflammation patterns, and the specific location of the inflammation(Amber, 2024; Le Berre et al., 2023). Even though CD is more severe than UC, its global prevalence is much lower (Gohil & Carramusa, 2014). Le Berre et al. (Le Berre et al., 2023) estimated that the number of people living with UC will reach 5 million worldwide by 2023. UC is a chronic condition characterized by colon and rectum inflammation, and the exact causes are unclear. However, according to recent studies (Basha et al., 2022; Clooney et al., 2019; Federici et al., 2022; Kennedy et al., 2024; Mills et al., 2022), UC is linked to an imbalance in the gut microbiota and gut bacteriophages. Patients with a long-standing history of UC and CD will have an increased risk of developing colorectal cancer (CRC) (Sato et al., 2023). Colorectal cancer (CRC) is the fourth most frequently occurring cancer in the United States and the second most common in Europe, causing the second most cancer deaths after lung cancer (*Colorectal Cancer Statistics | CDC*, 2022; *Colorectal_cancer_factsheet-Mar_2021.Pdf*, n.d.). In 2023, an estimated 153,020 individuals will be diagnosed with CRC, with 52,550 will die due to the disease (Siegel et al., 2023). Although screenings and newer technologies are available, prevalence and mortality rates are still high, even in high-income countries (Schmitt & Greten, 2021). Colorectal cancer can be caused by several factors, including genetic alterations, general factors such as climate, stress, and education, external influences like alcohol, smoking, and radiation, as well as internal factors including metabolism, inflammation, and the gut microbiome (Rebersek, 2021; Schmitt & Greten, 2021; Siegel et al., 2020).

1.6 Objective of this study

In this work, I first developed a tool, RepliDec, for predicting the phage replication cycle (temperate, virulent, and chronic). This tool can be used to predict the lifecycle for the complete phage genome and fragment phage sequences assembled from metagenomic or virome data. Then, an integrated pipeline, RepliDec+, was also created to complement RepliDec for complex environments, such as a soil sample. I used RepliDec+ to predict the phage lifecycle in inflammatory bowel disease (IBD) patients and healthy controls to get an overview of the phage's impact on IBD patients. In addition, I also used RepliDec to validate whether temperate phages exist in commercial cocktails.

To explore the impact of phage-bacteria interactions on IBD patients, 11 fecal samples were collected. Bacterial and viral resources from patients with UC, CRC, and healthy individuals were pooled together. Subsequently, we employed viral tagging on the cross-over infection samples to link bacteria with virome sources, revealing the dynamics of phage-bacteria interactions. This method enhances our understanding of the interactions between bacteria and their phages in complex natural settings, particularly during the development or progression of UC and CRC.

2. Material and Methods

2.1 Development of phage lifecycle prediction tool

2.1.1 RepliDec training dataset creation (Viral_Protein_DB)

I have created an in-house database called Viral_Protein_DB by obtaining viral and bacterial genomes from the NCBI RefSeq Database (O'Leary et al., 2016) (Appendix Table 1, Appendix Table 2).

Each viral genome was manually curated and classified as a prokaryotic virus based on host information, taxonomy, and organism name (Appendix Table 1; see 'Manual curation' section below for details). Following the curation steps, I could identify 4,596 unique prokaryotic viral genomes. Out of these, 4,126 viral genomes were used for further analysis, while 470 were reserved for test dataset 1 since their lifecycles were experimentally validated (Appendix Table 3). The amnio acid sequences from viral genomes were predicted using Prodigal v2.6.3 (Hyatt et al., 2010). I utilized BACPHLIP (Hockenberry & Wilke, 2021) to forecast the replication cycles of these viral genomes.

Bacterial genomes were acquired from NCBI on October 27, 2021, using the 'RefSeq' category set to 'reference genome' or 'representative genome' (Appendix Table 2). Prophage sequences were extracted from a total of 14,922 bacterial and archaea genomes using VIBRANT v1.2.1 (Kieft et al., 2020), which utilizes neural networks of protein signatures and a unique v-score metric to maximize the identification of lytic viral genomes and integrated proviruses. Prophage proteins were predicted using VIBRANT (v1.2.1) (Kieft et al., 2020). Proteins from prophages with lengths greater than 80 amino acids were retained. The replication cycles of predicted prophages are recognized as 'Temperate'.

In total, 419,983 proteins were found in 4,126 viral genomes, and 739,196 were found in 21,134 prophages. I clustered them using MMseqs2 (Steinegger & Söding, 2017) (Version: 7aade9df7475ae7c699b2074b5e4daa52e0245f1) with parameter "--cov-mode 0 --min-seq-id 0.70 -c 0.70") (Steinegger & Söding, 2017). 711,880 protein clusters (PCs) were generated, which served as the basis for calculating the likelihood required for Bayesian inference.

The tendency value of each protein cluster was calculated by $\log P(X|C_{Temperate})$ divided by $\log P(X|C_{Virulent})$ where X indicate the protein cluster. If the tendency value is greater than zero, the protein cluster tends to be lysogenic and vice versa. Tendency values range from -5.70 to 4.74 for all PCs. I identified high-tendency protein clusters from this range, which must contain at least 100 proteins with a tendency value exceeding 4 for "Temperate" or below -3.5 for "Virulent".

I annotated all amino acid sequences of all proteins using HMM searches against four databases: Pfam v34 (Mistry et al., 2021), KEGG (download date: 2022-02-01) (Kanehisa & Goto, 2000), VOGDB (211, https://vogdb.csb.univie.ac.at/), PHROGs (http://millardlab.org/2021/11/21/phage-annotation-with-phrogs/) (Terzian et al., 2021) under a threshold of 1e-5 using HMMER (v3.3.2) (Mistry et al., 2013). To determine the function of a PC, the most frequently occurring annotations within a cluster were chosen as the function of the cluster. I could annotate 80% of the proteins in the database (930,916) (Appendix Table 5).

Using Cytoscape (v3.9.1) (Shannon et al., 2003), I created a gene-sharing network of all genomes and protein clusters (PCs) based on their tendency and function.

2.1.2 Manual curation of viral genomes

To ensure the integrity of the training dataset utilized within RepliDec, I undertook a manual curation process to classify each viral genome, totaling 14,717, according to its type - whether eukaryotic or prokaryotic. The identification of prokaryotic viruses was conducted adhering to three predefined criteria:

1. Genomes containing host information in metadata will be classified as prokaryotic viruses if the host belongs to bacteria or archaea. 4408 genomes meet this criterion.

2. In my analysis of genomes with host information in metadata that does not fulfill the first criteria, I took into account the taxonomy and organism name. If the host is eukaryotic and the taxonomy includes '*Caudiviricetes*' or 'phage' in the organism name, each genome undergoes a manual verification to confirm its type. For example, NC_055902, identified as CrAssphage cr131_1, falls under '*Caudiviricetes*', was sourced from fecal samples, and the host is *Macaca mulatta*. Following this description and manual review, this genome will be classified as prokaryotic viruses despite potentially being categorized as eukaryotic viruses if the host data is solely considered.

3. When genome metadata lacks host details, I rely on taxonomy, organism names, and insights from the research paper. If the taxonomy is '*Caudiviricetes*' and the organism name includes 'bacteria', 'archaea', or 'phage', I verify these genomes by examining information in the research papers. For instance, NC_047700 - NC_047711 are uncultured phages categorized under '*Caudiviricetes*', derived from metagenomic data with unspecified hosts. While these genomes cannot be classified solely based on host information, the original article (Mizuno et al., 2013) indicates that they should be regarded as prokaryotic viruses.

Based on the three rules, I identified 4,596 prokaryotic viral genomes. Furthermore, 470 genomes were excluded from the training dataset and used as test dataset 1 (Appendix Table 3). In summary, 4,126 viral genomes were selected for establishing Viral_Protein_DB (Appendix Table 5).

2.1.3 Mathematic model of RepliDec

The Naïve Bayes Classifier technique (Rish, 2001) is based on the Bayesian theorem and is particularly suited for high dimensional inputs. To handle the complexity of high dimensionality, the Naïve Bayes classifier assumes that the features are independent of each other given a class, and one feature is not affected by others.

$$P(\mathbf{X}|C_k) = \Pi_{i=1}^{n} P(X_i|C_k)$$

Where X = $(X_1, \ldots, X_i)$ means the feature vector and C is class. Here, each PC is the feature vector and C is the life cycle type (k = (0,1), Virulent (C_1) or Temperate (C_0)). Based on the Bayesian theorem, it will be easy to calculate $P(C_k|\mathbf{X})$.

$$P(C_k|\mathbf{X}) = \frac{P(C_k)P(\mathbf{X}|C_k)}{P(\mathbf{X})}$$

Because the denominator is a constant, the formulation can be expressed as follows:

$$P(C_k|\mathbf{X}) \propto P(C_k)P(\mathbf{X}|C_k)$$

By using the chain rule, the final formulation is,

$$P(C_k|\mathbf{X}) \propto P(C_k)\Pi_{i=1}^{n} P(X_i|C_k)$$

Where X = $(X_1, \ldots, X_i)$ means the feature vector (PC) and C_k means two types of the replication cycle. $P(C_k)$ is the prior probability and $P(C_0)$ is the number of temperate phage genomes divided by the total number of genomes in Viral_Protein_DB. $P(C_1)$ is the number of virulent phages divided by the total number of genomes in Viral_Protein_DB.

To avoid $P(X_i|C_k)$ equal to 0, a small number α was added to each feature vector ($\alpha = 1$) given C_k , and to reduce the computation complexity, base 10 log form was applied.

$$\log P(C_k|\mathbf{X}) \propto \log P(C_k) + \log \Pi_{i=1}^{n} P(X_i|C_k)$$

$$\propto \log P(C_k) + \log \Pi_{i=1}^{n} \frac{n_i(C_k) + \alpha}{n_i + \lambda\alpha}$$

Where $n_i(C_k)$ denotes the number of the ith feature vector in the k class and λ denotes the number of class (here $\lambda = 2$).

If the $\log P\big(C_0\big|(X_1, \ldots, X_j)\big)$ > $\log P\big(C_1\big|X_1, \ldots, X_j\big)$ (which $j \leq n$), the replication cycle will be labeled as "Temperate" or "Virulent". Based on the mathematical model, I generated a probability profile for each PC given two replication cycles.

2.1.4 Test dataset preparation

To systematically compare the performance of all tools, I use two test datasets: (a) an experimentally established benchmark dataset (Dataset 1) and (b) a novel representative dataset (Dataset 2) under both complete genomes and metagenomic assemblies conditions.

Test dataset 1 was initially curated by Mavrich et al. (Mavrich & Hatfull, 2017) and used as the training dataset or test dataset in PhagePred, BAPHLIP, DeePhage, and PhaBox/PhaTYP. To minimize errors in the evaluation process, I refined the selection to 470 genomes. This ensured the inclusion only of those with experimentally validated results that align with bioinformatic predictions. In addition, during a manual review of these 470 genomes, two filamentous phages were found: *salmonella phage IKe* (NC_002014) and *pseudomonas phage Pf3* (NC_001418). It is well-documented that filamentous phages are extruded from the host cell via a process that does not result in the death of the host (Hay & Lithgow, 2019). Thus, the replication cycles of these two phages have been categorized as chronic rather than virulent. Test dataset 1 contains 207 temperate, 261 virulent, and two chronic.

Test dataset 2 includes 610 representative NCBI phage genomes, which can represent 2920 genomes, with no overlap with RefSeq and dataset 1 (at most 95% similarity). I generate this test dataset following these steps. I retrieved (date: 20230621) prokaryotic viral genomes from NCBI Virus with a "Sequence Type" not equal to "RefSeq" to avoid overfitting issues. Next, I used two steps to remove the redundancy of these non-RefSeq viral genomes. First, I used PSI-CD-HIT (v4.8.1) (W. Li & Godzik, 2006) with parameter "-c 0.98 -G 1 -g 1 -aS 0.7 -prog blastn" to remove redundancy within the non-RefSeq genomes based on a 70% coverage and 98% identity criteria. Next, I eliminate redundant representative genomes from the previous step by comparing them with those in the RepliDec training dataset, using average nucleotide identity (ANI) computed by FastANI (v1.33) (Jain et al., 2018). The longest genome was designated as the representative of the cluster. Representative genomes from singleton clusters (containing only one genome) were removed, as they are unique and not suitable for use as standard test dataset genomes. In total, 744 representative genomes were selected (cluster size ≥ 2).

I manually checked the metadata for each of these representative genomes to confirm that they are prokaryotic viruses. Unfortunately, 95 genomes belonging to the marine virus AFVG (e.g., MN694719.1; MN694804.1; MN694809.1; MN694810.1, and so on) do not have host or taxonomy information. There is no other information indicating that they are bacteriophages, so they were removed. The remaining genomes were refined to eliminate genomes with inconsistencies in lifecycle predictions (Appendix Table 3). After curation, 610 representative NCBI phage genomes were retained.

Test Dataset 1 and test Dataset 2 were predicted using PHACTS, BACPHLIP, PhaBOX/PhaTYP, DeePhage, PhageAI, and RepliDec to predict the replication cycle of these genomes.

2.1.5 Predict benchmark lifecycle for test dataset 2

To determine the replication cycle of these representative genomes, I employed two methods to verify it. Firstly, I use the same strategy developed by Mavrich et al. (Mavrich & Hatfull, 2017). I downloaded the conserved domain from the CDD database (available at 20230825) (M. Yang et al., 2020) and manually identified all conserved domains in the database containing descriptions related to "integrase" or "parA." Then, I predicted the amino acid sequences of each representative genome using Prodigal (v2.6.3; -g 11) (Hyatt et al., 2010), and checked if the proteins contained "integrase" or "parA" domain using RPS-BLAST (Reverse PSI-BLAST) (v2.13.0) (M. Yang et al., 2020) with an e-value of 1e-5. If a genome contains one of these "temperate" domains, it will be labeled as temperate; otherwise, it will be labeled as virulent. In addition, this method has an error rate of about 4%, according to the Mavrich article. So, I need to incorporate another method to confirm the replication cycle by aligning the representative genomes to bacteria genomes. I aligned these genomes to the NT (Sayers et al., 2022) prokaryotic database using BLASTn (v2.13.0) (M. Johnson et al., 2008) to identify temperate phages based on the prophage region. If a genome query coverage is greater than or equal to 50% and the corresponding identity is greater than or equal to 70%, then the genome will be labeled as temperate; otherwise, it will be labeled as virulent. If one of the above methods labels the genome as temperate, the final replication cycle will be considered temperate. After manually comparing the results from the two methods, I found that 39 genomes can be mapped to bacteria genomes with high query coverage and identity (e.g., the highest reach was 56% query coverage and 87% identity), which did not meet the criteria to be labeled as temperate. Genomes with inconsistent results were removed to ensure the test dataset's accuracy and reliability. If a representative genome belongs to the *Inoviridae* family, I labeled its replication cycle as chronic (Appendix Table 3).

2.1.6 Simulated test dataset

To evaluate the performance of these tools in assembled contigs from metagenomics or virome data. I simulated two sets of 150bp paired-end (PE) metagenomics reads from test dataset 1 and test dataset 2 using ART v2.5.8 (Huang et al., 2012) (parameter: -ss HS25 -p -l 150 -f 10 -m 200 -s 10). Then, I used fastp (S. Chen et al., 2018) to filter low-quality reads from simulated datasets with default parameters. Next, I assembled these reads into contig using SPAdes v3.15.2 (--meta) (Nurk et al., 2017), and only contigs longer than 3kb were kept for further analysis. To determine the replication cycle for these contigs, I aligned them back to the original genomes using Minimap2 (v2.1) (H. Li, 2018), and the replication cycle of a contig was assigned based on the most similar genomes.

I used PHACTS (McNair et al., 2012), BACPHLIP (Hockenberry & Wilke, 2021), PhaBOX/PhaTYP (Shang et al., 2023), DeePhage (Wu et al., 2021) , and RepliDec to predict the replication cycle of the assembled contigs from the two simulated test datasets.

2.1.7 Evaluation metrics

I used Sensitivity (Sn), Accuracy (Acc), F1-score, and the Matthews correlation coefficient (MCC) to evaluate the performance of different prediction tools, which are calculated as follows:

$$Sn = \frac{TP}{TP + FN}$$

$$\overline{Sn} = \left(Sn_{Temperate} + Sn_{Virulent} + Sn_{Chronic}\right)/3$$

$$Acc = \frac{TP + TN}{TP + TF + FP + FN}$$

$$= \frac{TP_{Temperate} + TN_{Temperate} + TP_{Virulent} + TN_{Virulent} + TP_{chronic} + TN_{Chronic}}{N_{Temperate} + N_{Virulent} + N_{Chronic}}$$

$$F1 - score = \frac{2 * Pre * Sn}{Pre + Sn}$$

$$\overline{F1 - score} = \left(F1 - score_{Temperate} + F1 - score_{Virulent} + F1 - score_{Chronic}\right)/3$$

$$MCC = \frac{c \times s - \sum_{k}^{K} p_k \times t_k}{\sqrt{(s^2 - \sum_{k}^{K} p_k^2) \times (s^2 - \sum_{k}^{K} t_k^2)}}$$

Where TP, FN, TN, and FP denote the numbers of true positive, false negative, true negative, and false positive, respectively. N indicates the number of cases.

For multiclass MCC, $t_k = \sum_{i}^{K} C_{ik}$ is the number of times class k truly occurred, $p_k = \sum_{i}^{K} C_{ki}$ is the number of times class k was predicted, $c = \sum_{k}^{K} C_{kk}$ is the total number of samples correctly predicted, $s = \sum_{i}^{K} \sum_{j}^{K} C_{ij}$ is the total number of samples

All the metrics were calculated using the sci-kit (v1.5.2) package (Pedregosa et al., 2011) in Python. The average sensitivity and average F1-score of multiclass are calculated using the “macro” parameter, which means calculating metrics for each lifecycle and finding their unweighted mean.

2.1.8 RepliDec+ pipeline

I have created an integrated pipeline that encompasses BACPHLIP, DeePhage, PhaBOX/PhaTYP, and RepliDec to manage complex microbiome environments. BACPHLIP is suitable for complete genomes, and DeePhage and PhaBOX/PhaTYP are suitable for incomplete contigs. I developed a new script to enable DeePhage and PhaBOX/PhaTYP handle multiple genomes as one result for binning. I designed a scoring system based on the evaluation results of these four tools from two test datasets. This scoring system can assign weights to each prediction result and output a final reliable prediction (Figure 2-1).

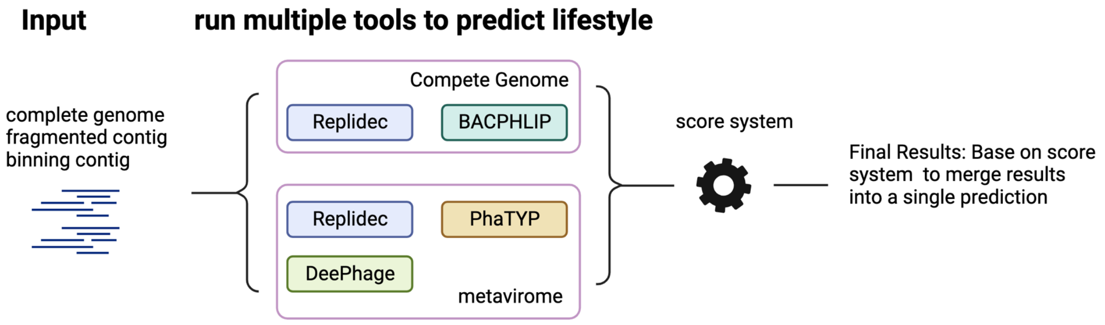


Figure 2-1 **RepliDec+ pipeline.**
RepliDec+ merges multiple tools based on a scoring system to provide an reliable replication cycle for viral sequences.

2.1.9 Detect temperate phages abundance in IBD patients using RepliDec+

I use an assemble-free strategy to assess viral abundance in patients with inflammatory bowel disease (IBD) using the Gut Phage database (GPD) (Camarillo-Guerrero et al., 2021) as a reference. Using RepliDec+, I predict the replication cycle of 142k viral genomes from the GPD dataset.

The raw reads from IBD patients and healthy controls were downloaded (Appendix Table 4) (Norman et al., 2015). Using fastp (v0.23.2) (S. Chen et al., 2018) to remove any low-quality reads and adapter sequences with the following parameters: "-z 4 -n 10 -l 60 -5 -3 -W 4 -M 20 -c -g -x". Then, the clean reads were mapped to the GPD database using Bowtie2 (v2.3.5.1) (Langmead & Salzberg, 2012) with the “--sensitive-local" parameter. We assess the relative abundance of viral sequences across various replication cycles by measuring it in transcripts per million (TPM).

2.2 Detect temperate phages in commercial cocktails

2.2.1 Commercial cocktails used in this study

Public data from four previously analyzed cocktails (PYO97, PYO2000, PYO2014, INTESTI) were downloaded from NCBI (ERR2184199, ERR2184200, ERR2184201, SRR3744915) (O'Leary et al., 2016). These cocktails were sourced from the Eliava Institute in Georgia and sequenced as previously described (Villarroel et al., 2017; Zschach et al., 2015). The published raw reads were used in this study.

2.2.2 Quality control and assembly

First, PhiX sequence was removed from the raw reads of all cocktails using Bowtie2 (v2.3.5.1)). PhiX sequences were identified in the following samples: PYO97 (8.43%), PYO2000 (23.07%), and PYO2014 (2.81%). Then, fastp (v0.23.2) (S. Chen et al., 2018) was used to control the read quality before the clean reads were assembled into contigs using SPAdes (--meta, v3.15.2) (Nurk et al., 2017). Contigs shorter than 1 kb were removed. CheckV (v0.8.1) (Nayfach et al., 2021) was used to remove the host region and assess the quality of the viral contigs. VirSorter2 (Guo et al., 2021) was also used after CheckV to identify the viral contig. For the cocktail sequences, two criteria were applied to ensure the assembled contig is viral: the contig quality is not equal to "not-determined" as assessed by CheckV, and the contig is classified as a virus by VirSorter2. Contigs that met these two criteria were used for future characterization analysis.

2.2.3 Relative abundance calculation

To estimate the viral contig's abundance within the cocktails, clean reads were mapped to the assembled contigs using Bowtie2 (v2.3.5.1). The number of mapped reads was calculated using SAMtools (v1.13) (Danecek et al., 2021). The relative abundance was calculated using the following formula:

$$\frac{R_x * 100}{L_x * \sum_n \frac{R_i}{L_i}}$$

In which R_x denotes the number of reads mapped to a contig, L_x is the length of the contig, and $\sum_n \frac{R_i}{L_i}$ corresponds to the sum of mapped reads (R_i) normalised by contig length (L_i).

2.2.4 Assessment of bacterial contamination

To determine any bacterial contamination in the sequences, unmapped reads were collected after mapping to the contigs, which excluded host regions using CheckV (v0.8.1) (Nayfach et al., 2021). These unmapped reads were then classified with Kraken2 (v2.1.2) (Wood et al., 2019) using the MinusB (v 12_9_2022) database, providing an overview of bacterial contamination. Bacterial genomes with the highest number of mapped reads were retrieved from NCBI (O'Leary et al., 2016) to validate the mapping regions alongside unmapped reads, employing Bowtie2 (v2.3.5.1). The genomes of *Enterococcus faecium* (GCF_009734005.1), *Escherichia coli* (GCF_000005845.2, GCF_000008865.2), *Proteus mirabilis* (GCF_000069965.1), *Shigella flexneri* (GCF_000006925.2), *Pseudomonas aeruginosa* (GCF_000006765.1), *Serratia marcescens* (GCF_003516165.1), and *Staphylococcus aureus* (GCF_000013425.1) were downloaded and examined, with unmapped reads mapped to each of these bacterial genomes using Bowtie2 (v2.3.5.1) with the "--sensitive-local" parameter.

2.2.5 *In Silico* characterization of phages

To detect if the temperate exists, RepliDec (v0.2.3.1) was used to predict the replication cycle of the viral contigs. The host information was predicted using iPHoP (v1.1.0) (Roux et al., 2022) with the default setting, and the predicted host with the highest score for each phage was assigned as the host. The taxonomy of viral contigs was assigned using the MMseqs2 taxonomy module (v13.45111) (Steinegger & Söding, 2017) with the Swiss-Prot database (release 2022_04) (Bairoch & Apweiler, 2000).

2.2.6 Phylogenetic analysis of the endolysins from the cocktail

I also investigate the endolysin diversity among these viral contigs. Endolysin genes were identified by annotating genes from all contigs with the PHROGs (v3) (Terzian et al., 2021) database using HMMER3 (v3.3.2). A phylogenetic tree was constructed using these endolysin proteins. The sequences of their amino acids were aligned using MAFFT (v7.505, mode: mafftlinsi) (Katoh et al., 2002) , and gaps were removed by trimAl (v1.4.rev15l, -gappyout) (Capella-Gutiérrez et al., 2009). The tree was constructed using IQ-TREE (v2.0.3) (Minh et al., 2020) with 1000 ultrafast bootstrap replications and the WAG+R4 substitution model, as recommended by the ModelFinder module in IQ-TREE. The tree was visualized using iTOL (Letunic & Bork, 2021).

2.3 Identify phage-bacteria pairs in patients with intestinal disorders

2.3.1 Sample collection and preparation

Stool samples were gathered from three ulcerative colitis patients, three colorectal cancer patients in the early stages (n=3), two in the late stages (n=2), and three healthy controls. My collaborator handled the isolation and DNA extraction from each sample to study the bacterial and viral communities. To obtain the viral-host relationship, the isolated bacteria and virome communities were pooled based on each condition and used for cross-infection studies (Table 2-1). My collaborator used a modified version of viral tagging, previously described in Unterer et al. (Unterer et al., 2023) , to sort the cells containing phage with their host attached to the surface. Following the manufacturer's guidelines, cells sorted from viral tagging, viral DNA, and bacterial DNA from subjects were amplified with the Repli-G kit (150343, Qiagen), and subsequently, sent for sequencing.

Table 2-1 List of all cross-infections samples

Healthy$_{\text{Bacteria}}$ + Healthy$_{\text{VLPs}}$	Ulcerative Colitis $_{\text{Bacteria}}$ + Healthy $_{\text{VLPs}}$	Early CRC $_{\text{Bacteria}}$ + Healthy $_{\text{VLPs}}$	Advanced CRC $_{\text{Bacteria}}$ + Healthy $_{\text{VLPs}}$
Healthy $_{\text{Bacteria}}$ + Ulcerative Colitis $_{\text{VLPs}}$	Ulcerative Colitis $_{\text{Bacteria}}$ + Ulcerative Colitis $_{\text{VLPs}}$	Early CRC $_{\text{Bacteria}}$ + Ulcerative Colitis $_{\text{VLPs}}$	Advanced CRC $_{\text{Bacteria}}$ + Ulcerative Colitis $_{\text{VLPs}}$
Healthy $_{\text{Bacteria}}$ + Early CRC$_{\text{VLPs}}$	Ulcerative Colitis $_{\text{Bacteria}}$ + Early CRC$_{\text{VLPs}}$	Early CRC $_{\text{Bacteria}}$ + Early CRC$_{\text{VLPs}}$	Advanced CRC Bacteria + Early CRC$_{\text{VLPs}}$
Healthy $_{\text{Bacteria}}$ + Advanced CRC $_{\text{VLPs}}$	Ulcerative Colitis $_{\text{Bacteria}}$ + Advanced CRC $_{\text{VLPs}}$	Early CRC $_{\text{Bacteria}}$ + Advanced CRC $_{\text{VLPs}}$	Advanced CRC $_{\text{Bacteria}}$ + Advanced CRC $_{\text{VLPs}}$

2.3.2 Native faecal metagenomes

Raw reads from fecal metagenomes (11 samples in total) were removed PhiX (reference: NC_001422.1) and human (reference: GRCh38) contamination using Bowtie2 (v2.3.5.1) (Langmead & Salzberg, 2012) with parameter "--sensitive-local" and SAMtools (v 1.17) (Danecek et al., 2021). Then low-quality reads were removed using fastp (v0.23.2) (S. Chen et al., 2018) with parameter "-z 4 -n 10 -l 60 -5 -3 -W 4 -M 20 -c -g -x". Clean reads from each sample were individually assembled using metaSPAdes (v3.15.2) (Nurk et al., 2017) with the default setting. Scaffolds length longer than 1kb were kept for further analysis.

To prevent the loss of shorter scaffolds, a supplementary public human gut database was utilized as a reference to enhance the de novo scaffolds. Clean reads from each sample were aligned with the Unified Human Gastrointestinal Genome (UHGG) catalog (v2.0.2) (Almeida et al., 2021) using Bowtie2 (v2.3.5.1) with the "--sensitive-local" parameter. Sequences that achieved at least 80% of the maximum mapping reads, more than 60% mapping coverage, or mapping depth greater than 80% of the maximum depth were preserved based on the outputs from SAMtools (v 1.17). Retained UHGG sequences and de novo assembled scaffolds were merged by each sample and removed redundancy at 95% similarity using CD-HIT (v4.8.1, psi-cd-hit) (W. Li & Godzik, 2006) with parameter "-c 0.95 -G 1 -g 1 -aL 0.7 -aS 0.7 -circle 1". Prophage regions of the non-redundancy representative sequences of each sample were annotated and identified using "annotate" and "find-proviruses" modules in geNomad (v1.7.4) (Camargo et al., 2024).

To create a non-redundant catalog of representative bacterial sequences from all samples (referred to as NRbacteria), we merged the sequences and removed redundancies to achieve a similarity threshold of 95% using CD-HIT (v4.8.1, psi-cd-hit) (W. Li & Godzik, 2006) with parameter "-c 0.95 -G 1 -g 1 -aL 0.7 -aS 0.7 -circle 1". Subsequently, the relative abundance was obtained by mapping the clean reads from each sample against to the NRbacteria with Bowtie2 (v2.3.5.1; configured with "--sensitive-local") and SAMtools (v1.17). The relative abundance of each sample was calculated by CoverM (v0.6.1) with the specified parameters "-m tpm covered_bases length" (B. Woodcroft, unpublished, https://github.com/wwood/CoverM). For the NRbacteria taxonomy assignment, the easy-taxonomy subcommand in MMseqs2 (v13.45111; DB: Swiss-Prot) and Kraken2 (v2.1.2; DB: MiniKraken_DB_8GB) (Wood & Salzberg, 2014) with the default setting were applied. MMseqs2 employs the lowest common ancestor (LCA) strategy, aligning with target amino acid sequences to predict taxonomy. However, Kraken2 utilizes exact-match database queries of k-mers for taxonomy prediction, representing a distinct methodological approach. CRISPR was identified from NRbacteria using CRISPRidentify (v1.2.1) (Mitrofanov et al., 2021) for viral host prediction in future analysis.

Additionally, a catalog of representative non-redundant prophage sequences from all samples (NRprophage) was created to evaluate the impact of prophages on NRbacteria and viral tagging samples. This process mirrors the development of NRbacteria, utilizing CD-HIT (v4.8.1, psi-cd-hit) with the parameters "-c 0.95 -G 1 -g 1 -aL 0.7 -aS 0.7 -circle 1”. Subsequently, the cleaned reads from each sample were mapped to NRprophage using Bowtie2 (v2.3.5.1, “--sensitive-local"), and the relative abundance was computed with CoverM (v0.6.1; parameter: “-m tpm covered_bases length”; B. Woodcroft, unpublished, https://github.com/wwood/CoverM).

2.3.3 Virome

Approximately 2GB of 2×150bp sequence data were obtained per sample. Raw reads from Virome (11 samples in total) were removed PhiX (reference: NC_001422.1) and human (reference: GRCh38) contamination using Bowtie2 (v2.3.5.1) with parameter “--sensitive-local” and SAMtools (v 1.17). Then low-quality reads were removed using fastp (v0.23.2) (S. Chen et al., 2018) with parameter “-z 4 -n 10 -l 60 -5 -3 -W 4 -M 20 -c -g -x”. Clean reads from each sample were individually assembled using metaSPAdes (v3.15.2) (Nurk et al., 2017) with the default setting. Scaffolds length longer than 1kb were kept for further analysis. Clean reads were mapped back to the assembled contigs to evaluate assemble performance using Bowtie2 (v2.3.5.1) with parameter “--sensitive-local” and SAMtools (v 1.17), and the average mapping rate was 91.59%.

To further identify viral sequences, VirSorter (v1.0.6) (Roux et al., 2015) was used to identify putative virion contigs (VirSorter categories 1, 2, and 3) using both database options: -db 1 (RefSeq viruses) and -db 2 (RefSeq viruses+viromes) after CheckV (v0.8.1; end-to-end) removed the bacteria region of assembled scaffolds.

A non-redundancy viral contigs catalog (NRVvirome) was created to assess viral diversity within the subjects. Viral contigs detected with VirSorter were subjected to a dereplication process at a 95% similarity threshold. Pairwise distances were computed using Mash (v2.3) with a k-mer size of 21 and a sketch size of 10000 (Ondov et al., 2019). The cluster was assigned using an in-house script (mash_clstr.py) based on Mash-distance, and the longest sequence was kept as the representative of each cluster. Clean reads were mapped to NRVvirome using Bowtie2 (v2.3.5.1) with parameter “--sensitive-local” and SAMtools (v 1.17). The mapping rate ranges from 80.38% to 1.95%. The low mapping rates are caused by only 5 viral contigs identified in the sample. The abundance was estimated using CoverM (v0.6.1; parameter: “-m tpm covered_bases length”; B. Woodcroft, unpublished, https://github.com/wwood/CoverM). Viral contigs were taxonomically assigned using the ‘annotate’ function in geNomad (v1.7.4) (Camargo et al., 2024). All were classified under the categories “*Caudoviricetes*” and “*Malgrandaviricetes*” based on the latest viral taxonomy rules (Turner et al., 2023). iPHoP (v1.1.0) (Roux et al., 2022) was used to predict the viral host.

2.3.4 Viral tagging

Approximately 2GB of 2×150bp sequence data were obtained per sample. Reads were quality trimmed using fastp (v0.23.2) with parameter “-z 4 -n 10 -l 60 -5 -3 -W 4 -M 20 -c -g -x”. Reads were assembled using metaSPAdes (v3.15.2) with the default setting. Scaffolds length longer than 1kb were kept for further analysis. Approximately 99% of clean reads were recruited in assembled contigs calculated using Bowtie2 (v2.3.5.1, parameter “--sensitive-local”) and SAMtools (v 1.17). Viral contigs were identified by CheckV (v0.8.1; end-to-end) and VirSorter (v1.0.6), following the same steps in virome analysis. A non-redundancy viral contigs catalog from viral tagging (NRVvt) samples was also created at a threshold of 95% similarity using Mash (v2.3) (Ondov et al., 2019) with a k-mer size of 21 and a sketch size of 10000. An in-house script (mash_clstr.py) assigned the clusters and their representatives, utilizing Mash-distance. The clean reads achieved their highest mapping rate to NRVvt at 8.14% using Bowtie2 (v2.3.5.1 with the "--sensitive-local" parameter) and SAMtools (v1.17) due to the limited identification of viral contigs.

2.3.5 Cross-assembly of virome and viral tagging

The cross-assembly technique was employed to reconstruct viral sequences, aiming to reduce errors linked to low mapping rates in NRVvirome and NRVvt. Virome samples were cross-assembled with viral tagging samples derived from the same disease group (H, UC, CRCE, and CRCA) using metaSPAdes (v3.15.2) with the default settings. Subsequently, the cross-assembled contigs were dereplicated by the disease group at a 95% similarity threshold using Mash (v2.3; parameter: -k 21 -s 10000) to conserve computational resources for future analysis. A representative for each cluster was identified through an in-house script (mash_clstr.py) based on Mash distance.

2.3.6 Identification of viral sequences and viral clusters from cross-assemblies

To compile a detailed catalog of viral sources, I utilized NRVvirome, NRVvt, and a public human gut phage database (GPD) as additional resources to cross-assemble contigs. Initially, I retained representative cross-assembled sequences that included at least one viral gene predicted by CheckV (v0.8.1; end-to-end). This resulted in the preservation of 4,084 viral sequences. Next, to address differences caused by various assembly techniques, I incorporated NRVvirome and NRVvt. I aligned viral sequences from NRVvirome and NRVvt separately to the cross-assembled contigs of four disease groups using BLASTn (v2.13.0) (M. Johnson et al., 2008) with parameter "-max_target_seqs 10". Cross-assembled contigs were retained if they met these criteria: (a) query coverage (qcov) was greater than 60, and (b) the percentage of identity (pct_identity) was greater than 85. 124 potential viral sequences were identified without any detected viral genes. Additionally, GPD was also used to expand the viral source in case some fragmented and low-depth contigs were missing and hard to assemble. Clean reads from virome samples and viral tagging samples were mapped to GPD contigs, respectively, using Bowtie2 (v2.3.5.1) with the parameter "--sensitive-local" and SAMtools (v 1.17). GPD sequences were recruited into the viral source if the reads coverage of a sequence greater than 60 or the cover base (covbases) was greater than 10k. 6676 GPD sequences were recruited from reads mapping.

A total of 10,884 viral sequences were analyzed, including 6,676 from GPD and 4,208 from cross-assembled contigs. I first eliminated any duplicate viral sequences we identified to establish the viral reference for analyzing the viral tagging samples. I obtained 7,336 viral clusters (NRVcross-assemble) using BLASTn (v2.13.0) (M. Johnson et al., 2008) and scripts from the CheckV repository (https://bitbucket.org/berkeleylab/checkv/src/master/). These clusters fulfill the 98% pairwise average nucleotide identity (ANI) and 85% minimum coverage criteria. A 98% pairwise ANI threshold was employed to maintain viral contigs at the strain level. I chose viral clusters instead of taxonomy details due to the recent phage taxonomy classification system. According to geNomad (v1.7.4) (Camargo et al., 2024), most viral sequences are categorized under "*Caudoviricetes*".

2.3.7 Host prediction of cross-assemblies

Host information was identified using CRISPR and computational predictive software. First, data for non-redundant contigs from four disease groups was gathered using the public CRISPR database, CrisprOpenDB (downloaded on 202404, with the default database available at: http://crispr.genome.ulaval.ca/dash/PhageHostIdentifier_DBfiles.zip) (Dion et al., 2021). These non-redundant contigs were also aligned to CRISPRs from NRbacteria using BLASTn (v2.13.0) with the parameter "-max_target_seqs 1000." We selected only those sequences with a percentage of identity (pct_identity) of 98 or higher and two or fewer mismatches (n_of_mismatches) as validated matches. If a CRISPR is detected from the bacteria, the bacteria's taxonomy is used as the host.

Host information of GPD and other contigs cannot be assigned using CRISPR; it was assigned using iPHoP (v1.1.0), with a confidence score greater than 95. All ambiguous host genus names, such as: "UMGS680", were manually removed.

2.3.8 Phage-bacteria association network

I investigated the presence of bacteria and phages in the crossover VT samples. Using Bowtie2 (v2.3.5.1 with the parameter "--sensitive-local") and SAMtools (v1.17), I mapped the clean reads of these samples to NRVcross-assemble and NRbacteria to identify the viral and bacterial sources separately. Sequences are considered validated if they contain at least 100 mapping reads and have coverage greater than 10.

Viral tagging reads can be mapped to 607 viral sequences in NRVcross-assemble. Additionally, I analyze the composition of these 607 viral clusters to ascertain whether the sequences are unique to a specific disease group, using the Python package UpSet (v0.9.0) (Lex et al., 2014) for visualization. Bacteria sources were chosen at the genus level to align with the predicted phage host from the computational tool. 208 bacteria genera were detected from VT cross-infection samples.

The reads count profile of 607 VCs and 208 bacteria genera was utilized to create association networks, which were calculated and visualized using the R package NetCoMi (v1.1.0) (Peschel et al., 2021). The centered log-ratio transformation (clr) normalization method was used, and "sparcc" was chosen as the measure of association. Network edges with a threshold below 0.95 were removed to keep the association network plot clean.

The read count profile is normalized utilizing Transcripts Per Million (TPM) as relative abundance, and visualization is conducted through the Python package "seaborn" (v0.13) (Waskom, 2021). Viral clusters were visualized as circular plots using BRIG (v0.95) (Alikhan et al., 2011) with default parameters. The innermost ring represents the longest and most complete reference contig.

2.3.9 Statistics

Due to the non-normally distributed characteristics of microbial data, appropriate statistical analyses were conducted using non-parametric tests. This included the Wilcoxon signed-rank test, implemented through the "wilcoxon" function in the stats module of the SciPy Python package (v1.14.1) (Virtanen et al., 2020).

2.4 Code availability

The RepliDec code is available at Github: https://github.com/deng-lab/RepliDec and https://github.com/pengSherryYel/RepliDec; The RepliDec+ code is available at GitHub: https://github.com/pengSherryYel/RepliDecPlus.

The codes for detecting temperate phages in commercial cocktail formulations are accessible on GitHub.: https://github.com/deng-lab/ProphageCocktail

The codes for identifying viral-host pairs using viral tagging are available on GitHub.: https://github.com/pengSherryYel/Codes_for_VT_CRC

3. Results

3.1 Lifecycle prediction tool: RepliDec and RepliDec+

This chapter contains the results of a manuscript that has been submitted to the bioRxiv preprint and submitted to the journal *GigaScience* and is currently under review:

Xue Peng, Mohammadali Khan Mirzaei, Jinlong Ru, Li Deng. 2024. RepliDec, a Naive Bayes classifier, and RepliDec+, an integrative framework for accurate phage replication cycle prediction in metagenomic data. GigaSicence. Under revision.

Remarks: X.P. developed the software, constructed the database, performed the analyses, and drafted the manuscript. J.R. provided inputs during the software development step. M.K.M. and L.D. conceived and supervised the project and revised the manuscript. All authors reviewed and approved the manuscript.

3.1.1 RepliDec pipeline

The RepliDec pipeline was written in Python, and some open-source software was used: Prodigal (-g 11) (Hyatt et al., 2010), HMMER3 (Mistry et al., 2013), MMseqs2 (Steinegger & Söding, 2017), and BLASTp (M. Johnson et al., 2008).

RepliDec follows 3 steps in its predictions (Figure 3-1):

Step 1: RepliDec identifies the two most prevalent genes found in temperate phages: integrase and excisionase. These two regulatory proteins control how the phage genome inserts into and excises from the host genome. I retrieved all Pfams associated with integrase and excisionase, along with those containing pertinent descriptions, from the PFAM Database (v4). If the input is the phage genome, proteins from the query genomes will be predicted using Prodigal; this step will be bypassed if the input is a protein dataset. Subsequently, RepliDec aligns the query proteins with the integrase (27 Pfam families) and excisionase (3 Pfam families) using HMMER3 (hmmsearch). If the query proteins include either of these two “temperate” markers with an e-value less than 1e-5, the query genome is classified as 'Temperate'; otherwise, it will be categorized as "Virulent".

Step 2: Computational replication cycle using the naïve Bayes Classifier. Not all temperate phage genomes possess integrase and excisionase, particularly those derived from metagenomics or virome data. Incomplete contigs can lead to the loss of some genes during the assembly process, making Step 1 less effective for such contigs. In this step, RepliDec employs proteins from RefSeq and prophages to mitigate the risk of missing marker proteins in fragmented contigs (greater than 3000 bp). RepliDec aligns the query protein with the in-house Viral_Protein_DB using MMseqs2 (easy-search) with the parameters '-s 7 --max-seqs 1 --alignment-mode 3 --alignment-output-mode 0 --min-aln-len 40 --cov-mode 0 --greedy-best-hits 1'. The calculation results in two types of replication cycles. Then, calculate the conditional probability of $P(C_k|\mathbf{X})$, Where **X** = $(X_1, \dots, X_i)$ means the mapped protein cluster and C_k means two types of replication cycles. If the $\log P(C_1|\mathbf{X})$ smaller than $\log P(C_0|\mathbf{X})\, so$, then the query genome will be labeled as 'Temperate' or, it will be 'Virulent'.

Step 3: Use HMMER3 (hmmsearch) and BLASTp to identify the chronic replication cycle based on PI-like proteins. Chronic replication cycles are found in filamentous phages, and PI-like proteins are highly conserved proteins despite the filamentous phage genome being highly diverse (Hay & Lithgow, 2019; Roux, Krupovic, et al., 2019). If proteins in the query include PI-like proteins with an e-value below 1e-6, the query genome is classified as "Chronic".

If step 1 or 2 gives a "Temperate" label, the final predicted result will be labeled "Temperate". If step 3 gives a "Chronic" label, the final predicted result will be labeled as "Chronic".

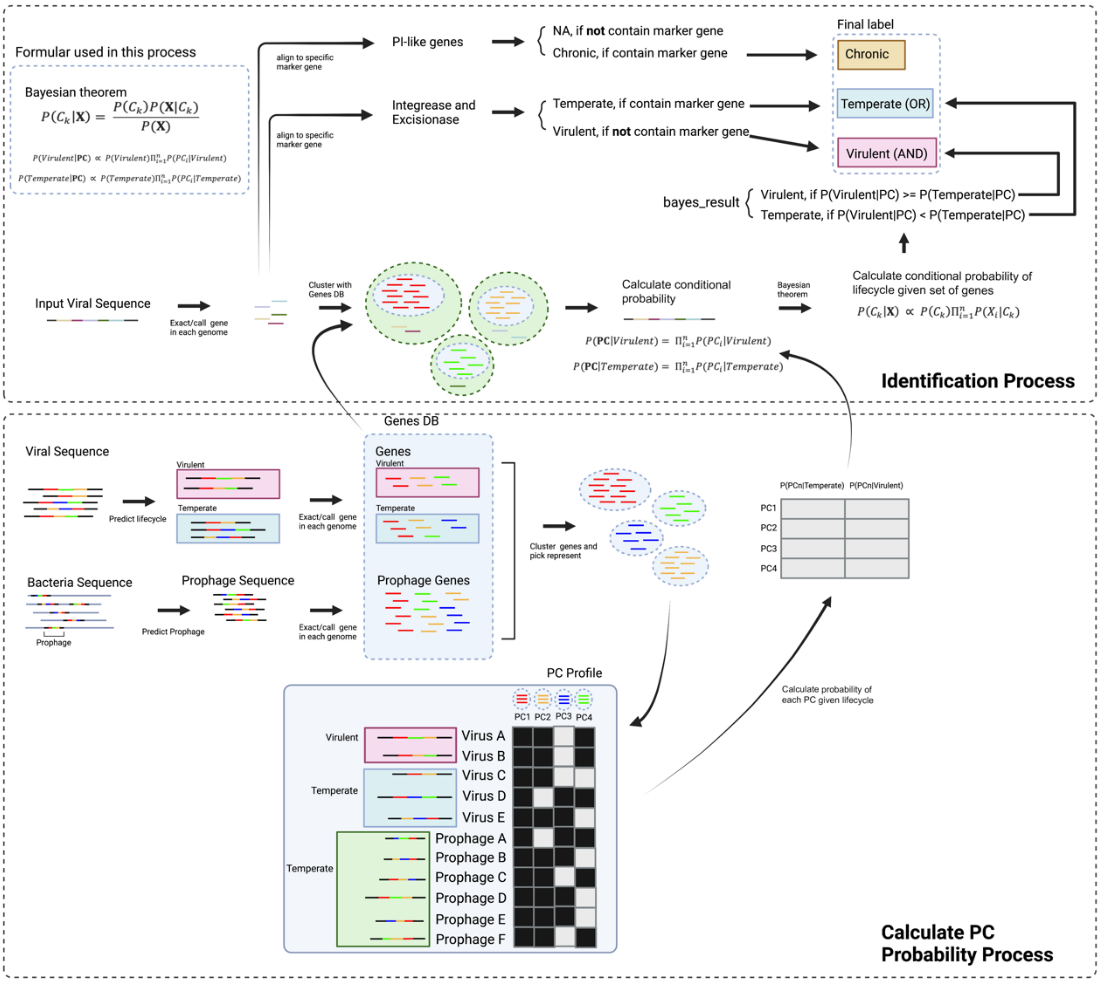

Formular used in this process
Bayesian theorem
$P(C_k|\mathbf{X}) = \frac{P(C_k)P(\mathbf{X}|C_k)}{P(\mathbf{X})}$
$P(Virulent|\mathbf{PC}) \propto P(Virulent)\Pi_{i=1}^{n}P(PC_i|Virulent)$
$P(Temperate|\mathbf{PC}) \propto P(Temperate)\Pi_{i=1}^{n}P(PC_i|Temperate)$
align to specific marker gene
PI-like genes
NA, if **not** contain marker gene
Chronic, if contain marker gene
align to specific marker gene
Integrease and Excisionase
Temperate, if contain marker gene
Virulent, if **not** contain marker gene
Final label
Chronic
Temperate (OR)
Virulent (AND)
bayes_result
Virulent, if P(Virulent|PC) >= P(Temperate|PC)
Temperate, if P(Virulent|PC) < P(Temperate|PC)
Input Viral Sequence
Exact/call gene in each genome
Cluster with Genes DB
Calculate conditional probability
$P(\mathbf{PC}|Virulent) = \Pi_{i=1}^{n}P(PC_i|Virulent)$
$P(\mathbf{PC}|Temperate) = \Pi_{i=1}^{n}P(PC_i|Temperate)$
Bayesian theorem
Calculate conditional probability of lifecycle given set of genes
$P(C_k|\mathbf{X}) \propto P(C_k)\Pi_{i=1}^{n}P(X_i|C_k)$
Identification Process
Genes DB
Viral Sequence
Predict lifecycle
Virulent
Temperate
Exact/call gene in each genome
Genes
Virulent
Temperate
Bacteria Sequence
Prophage
Predict Prophage
Prophage Sequence
Exact/call gene in each genome
Prophage Genes
Cluster genes and pick represent
P(PCn|Temperate)
P(PCn|Virulent)
PC1
PC2
PC3
PC4
Calculate probability of each PC given lifecycle
PC Profile
PC1
PC2
PC3
PC4
Virulent
Virus A
Virus B
Temperate
Virus C
Virus D
Virus E
Temperate
Prophage A
Prophage B
Prophage C
Prophage D
Prophage E
Prophage F
Calculate PC Probability Process

Figure 3-1 **RepliDec workflow for predicting the replication cycles of viral sequences.**

3.1.2 Biological features used in RepliDec

I developed a specialized viral protein database (Viral_Protein_DB) that serves as the training dataset for the computational analysis in RepliDec. This database includes 1,159,179 proteins, comprising 419,983 proteins from 4,126 prokaryotic viral genomes and 739,196 from 21,134 prophages identified across 14,922 bacterial and archaeal genomes. Notably, 930,916 proteins (about 80% in total) can be annotated in at least one of the four databases (Pfam v34, KEGG, VOGDB, PHROGs) at a significance level of 1e-5 using HMMER3 (Appendix Table 5). I listed 60 of the most frequent functions in this database, including 15 functions related to the lytic-lysogenic process (98,337 genes), which only account for 8.4% of genes in total (Figure 3-2 A). We detected some known lysogenic phage biomarkers (i.e., integrase, transposase, excisionase, resolvase, and recombinase) according to the previous studies (Muscatt et al., 2022; Tang et al., 2023) in our database. As anticipated, the most recognized “temperate” marker gene integrase (~18.6k proteins), which assists phages in integrating their genomes into host genomes, is the most frequently occurring protein in our database, aside from the structural protein (tail protein, 19k). Additionally, the database includes 6,308 transposases, 5,163 resolvases, 2,931 RusA-like holiday junction resolvases, 1,245 excisionases, 587 UvsX-like recombinases, and 414 recombinases.

The database also comprises proteins involved in the lytic-lysogenic switch system, including the CI-CII-Cro and parABS systems. It consists of 86 CI-like proteins, 1,891 CII-like transcriptional activators, and 25 Cro proteins. Additionally, 758 parA proteins and 1,695 parB proteins are associated with the parABS system.

Additionally, this database includes genes that not only regulate the lytic-lysogenic switch but also those associated with the lysis process, such as endolysin (16,358 proteins), holin (7,295 proteins), and Rz-like spanin (4,135 proteins). These genes mentioned above represent only 8.4% of the total genes in this database, indicating that a substantial number of genes still require verification. As experimental testing is unfeasible, we employed computational measurements to refine our search parameters (refer to the next section).

3.1.3 Tendency of Viral_Protein_DB protein cluster

I characterize each protein cluster's tendency based on its probability of presence in either the virulent or temperate phage genome. The tendency is determined by the logarithm of the lysogenic probability divided by the lytic probability. This enables me to assess the protein clusters often occurring in temperate phages. The tendency value ranges from -5.70 to 4.74. The criteria for defining a high tendency protein cluster are: a) the cluster must contain at least 100 proteins; b) a cluster with a tendency value greater than 4 is considered to have a high “Temperate” tendency; c) a cluster with a tendency value less than -3.5 is classified as having a high “Virulent” tendency. Integrase is a confirmed gene involved in the lysogenic-lysis process (Colavecchio et al., 2017; Smith & Jeddeloh, 2005), which exhibits a strong temperature tendency with values around 4.34 (Figure 3-2 B).

The protein cluster (PC_591446) has the highest temperate value (4.74) with a cluster size equal to 171. Its function remains unclear, yet the strong temperate tendency suggests it is prevalent in

171 temperate genomes. Additionally, there are six protein clusters that show a strong temperature tendency, but they lack any annotations. Unfortunately, there is currently no experimental evidence to support that they have the potential to serve as biomarkers to predict the lysogenic lifecycle.

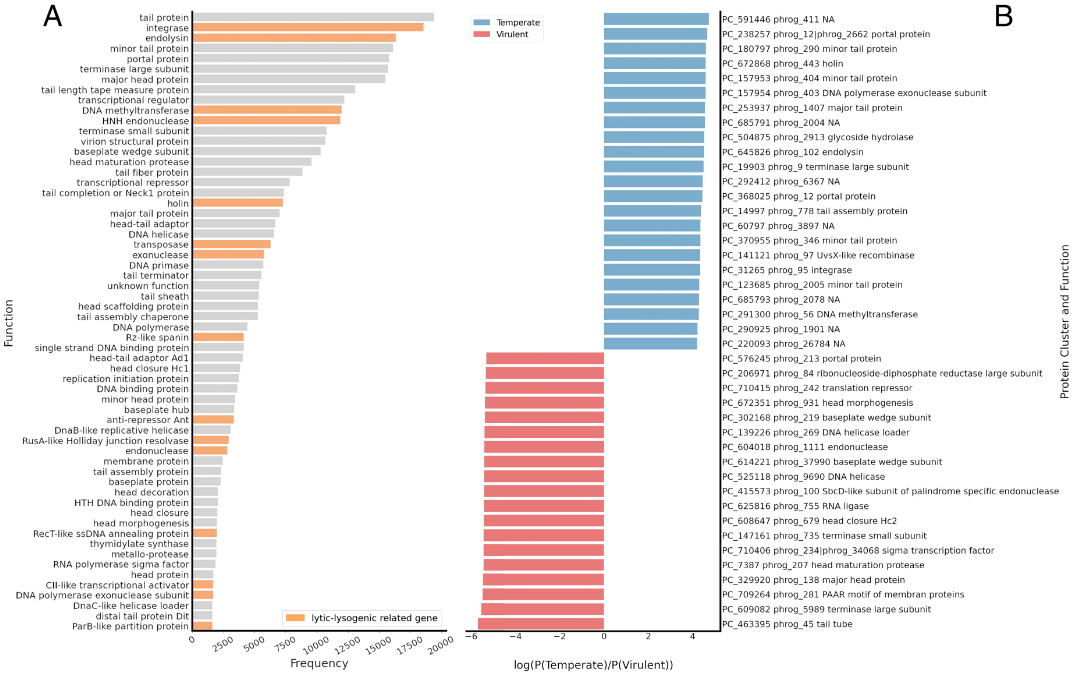


Figure 3-2 **Function annotation for Viral_Protein_DB proteins and tendency of protein cluster.**
(A) The distribution of 60 top frequency functions in Viral_Protein_DB; (B) High tendency protein cluster. Protein clusters have a minimum size of 100 and a tendency value greater than and equal to 4 or smaller than and equal to -5.3. The tendency value is defined as the probability of a protein cluster coming from temperate phages divided by the probability of the protein cluster coming from virulent phages in a log-transformed.

I established a gene-sharing network to examine the relationship between phages with varying replication cycles and their associated protein clusters. This gene network consists of one large sub-network, two medium sub-networks, and several small sub-networks (Figure 3-3 A). Temperate and virulent phages are distinctly separated within the large sub-network (Figure 3-3 B), and observing the patterns of each protein cluster reveals a clear boundary between the two phage types (Figure 3-3 C). Furthermore, protein clusters associated with integration and excision functions are predominantly located in the temperate phages within the large sub-network (Figure 3-3 B).

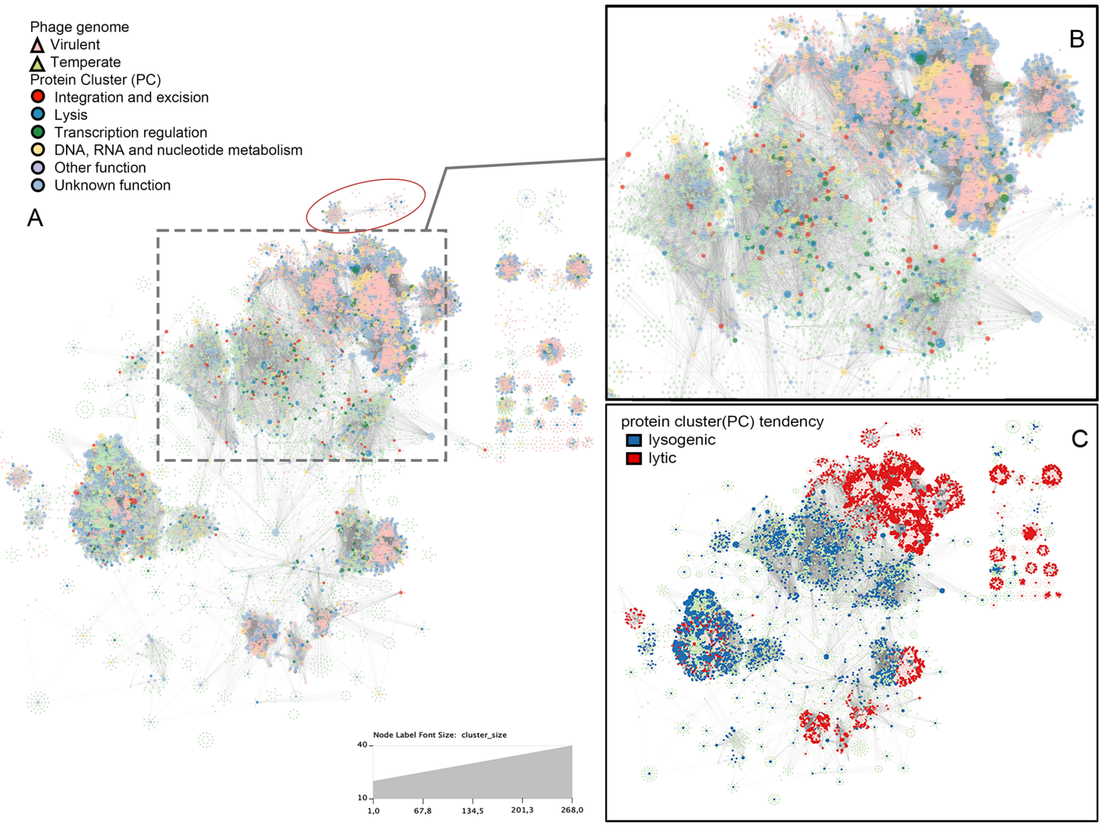


Figure 3-3 **The network between protein clusters and phages genomes.**
(A and B) a network of protein clusters showing different functions and B is approximately outlined as a dashed box in A and (C) a network of protein clusters showing different tendencies. Triangle nodes indicate phage genomes, with light green representing temperate phages and light pink representing virulent phages. The size of triangle nodes indicated how many protein clusters were connected to the phage. Round nodes indicate protein clusters. Round color nodes in A indicate the function of each PC and that in C indicate the tendency of each PC. Notes: protein clusters related to structure and DNA, RNA_and_nucleotide_metabolism are removed to decrease the figure's complexity.

3.1.4 RepliDec performance on experimentally validated dataset (Test Dataset 1)

Although RepliDec has the largest training dataset, I still need to evaluate its effectiveness in predicting replication cycles. I chose 470 RefSeq viral genomes that were experimentally validated and aligned with the bioinformatics predictions produced by Marvich (Mavrich & Hatfull, 2017), and this dataset has been used in PhagePred, BAPHLIP, DeePhage, and

PhaBox/PhaTYP. First, we compared the performance on complete genomes (Figure 3-4 A-D), RepDec excels among all tools, achieving high metrics: sensitivity (98.43%), accuracy (97.66%), F1 score (91.81%), and Matthews correlation coefficient (MCC, 95.31%). PhageAI also performs well across the same four measurements: sensitivity (95.71%), accuracy (93.40%), F1 score (95.55%), and MCC (86.91%). Conversely, DeePhage, BACPHLIP, and PhaBox/PhaTYP demonstrate high accuracy and MCC scores but struggle with sensitivity and F1 scores. For instance, DeePhage reaches 94.90% accuracy, while both BACPHLIP and PhaBox/PhaTYP approach near-perfect accuracy (BACPHLIP: 99.36%; PhaBox/PhaTYP: 99.15%) but exhibit low F1 scores (DeePhage: 63.39%; BACPHLIP: 66.40%; PhaBox/PhaTYP: 66.25%). This is attributed to their inability to predict the chronic replication cycle, which lowers the unweighted mean of the F1 score and sensitivity when calculating multiclass targets. Among all tools, PHACTS ranks the lowest in the four measurements.

I also assessed the performance on fragmented contigs simulated by complete phage genomes from Dataset 1 to mimic metagenomic data (Figure 3-4 E-H). PhageAI was not considered in the simulated data due to the 100 query/day limitation, which will take a very long time to predict thousands of assembled contigs. RepliDec performs impressively on assembled contigs, achieving 96.24% accuracy and 90.82% MCC, slightly lower than PhaBox/PhaTYP, which boasts 97.90% accuracy and 94.70% MCC. Additionally, RepliDec excels in other metrics, scoring 75.82% in F1 score and 97.47% in MCC. DeePhage, designed for virome data, also demonstrates strong performance across all four metrics, with sensitivity at 64.04%, accuracy at 94.57%, F1 score at 62.38%, and MCC at 87.61%. It accurately predicts most temperate contigs, though it labeled 107 as “Temperate” and incorrectly categorized 2 from chronic phages as “Virulent.” DeePhage tends to classify contigs as temperate, unlike BACPHLIP, which tends to label temperate phages as virulent. For instance, BACPHLIP misclassified 467 contigs from temperate phages as “Virulent” and made similar errors with chronic phages (Appendix Figure 1). Overall, RepliDec and phageAI excel in complete genome predictions, particularly for three types of replication cycles, and RepliDec also shows commendable performance in predicting fragmented contigs.

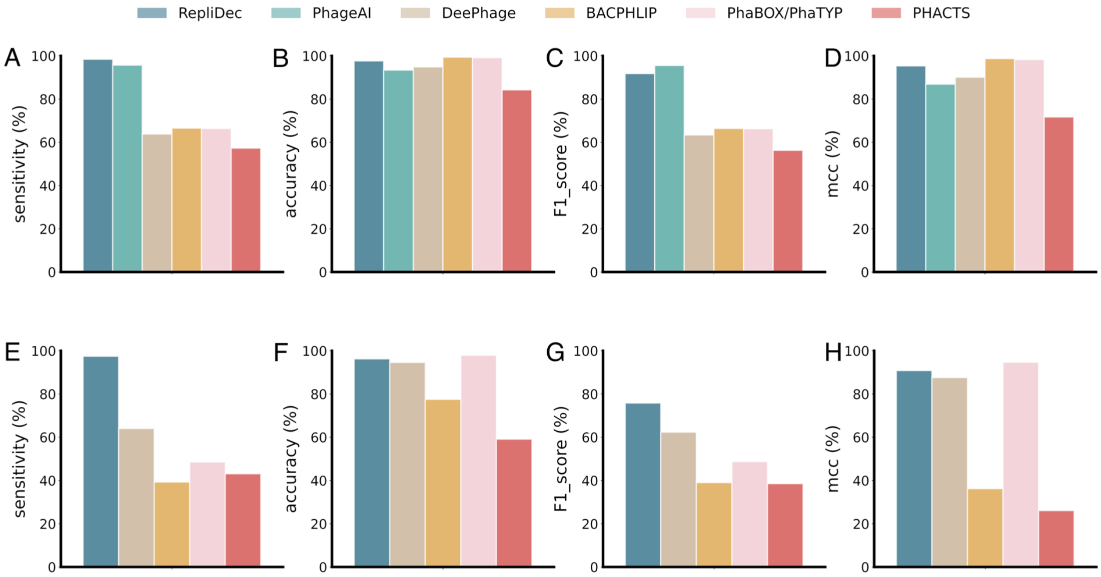


Figure 3-4 **Performance of different tools in predicting the lifecycle of experimental checked test dataset (Test Dataset 1).**

Comparison of complete phages genomes between different tools on sensitivity(A), accuracy(B), f1_score(C), MCC(D). Comparison of assembled contig simulated from test dataset 1 genomes between different tools on sensitivity(E), accuracy(F), f1_score(G), MCC(H).

3.1.5 RepliDec performance on bioinformatic predicted test dataset (Test Dataset 2)

RepliDec demonstrates strong performance on test dataset 1, even though these genomes were not part of the training set. To prevent overfitting and ensure valid results, we created a novel dataset, considering that some genomes from the RepliDec training set may have been similar (over 98% similarity) to those in test dataset 1. To evaluate performance across all tools systematically and fairly, this new benchmark dataset excludes genomes from the training datasets of these tools. It comprises 610 representative genomes from a total of 2,920, ensuring each exhibits no more than 95% similarity to the others and avoids similarities with RefSeq or prophage sequences. I initially evaluated the performance of complete genomes. RepliDec achieved top scores across four metrics: sensitivity at 86.76%, accuracy at 85.57%, F1 score at 87.31%, and MCC at 74.00%. BACPHLIP came in second with sensitivity at 59.55%, accuracy at 83.77%, F1 score at 57.21%, and MCC at 71.69% (Figure 3-5 A-D). Interestingly, PhageAI, DeePhage, and PhaBox/PhaTYP performed worse on this new dataset than on test dataset 1, indicating that these tools struggle with accurately predicting remote genomes, particularly novel phages with low similarity (95% similarity) to RefSeq, which is their training dataset.

I evaluated the performance of these tools on simulated assemblies (Figure 3-5 E-H). A total of 2,152 contigs, each with a minimum length of 3 kb, were assembled from test dataset 2, ranging from 3,860 to 387,980 bp with an average of 44,446.9 bp. All tools experienced a slight decline in four measurements when predicting the replication cycle of fragmented contigs compared to the complete genome. This indicates that the input length significantly affects prediction results. RepliDec outperformed all tools across all measurements (sensitivity: 66.74%, accuracy: 77.04%, F1 score: 70.69%, MCC: 55.87%). RepliDec accurately classified 1,184 (96.18%) temperate contigs, 454 (56.29%) virulent contigs, and 20 (52.63%) other contigs (Appendix Figure 2). DeePhage and PhaBox/PhaTYP ranked second and third in accuracy, respectively (DeePhage: 70.26%; PhaBOX/PhaTYP: 61.94%), F1 score (DeePhage: 47.24%; PhaBOX/PhaTYP: 32.88%), and MCC (DeePhage: 43.54%; PhaBOX/PhaTYP: 34.23%). DeePhage correctly predicted 795 (64.58%) temperate contigs and 717 (81.20%) virulent contigs, surpassing PhaBOX/PhaTYP, which predicted 720 temperate contigs and 613 virulent contigs (Appendix Figure 2 A and C). This suggests that DeePhage and PhaBox/PhaTYP excel at predicting virulent contigs, while RepliDec demonstrates high accuracy in predicting temperate contigs.

DeePhage and PhaBox/PhaTYP handle chronic phage contigs in distinct ways; importantly, DeePhage predicts most chronic contigs as temperate, while PhaBox/PhaTYP categorizes them as virulent. This discrepancy suggests a frequent misclassification of chronic contigs across various tools, complicating the analysis of metagenomic and virome data. Given the widespread presence of chronic phages in bacteria and archaea (Roux, Krupovic, et al., 2019), neglecting them in metagenomic and virome data analysis is unwise, particularly in complex metagenomic assemblies.

Additionally, BACPHLIP showed a significant decrease in performance compared to complete genomes, likely because it is primarily designed for those genomes. The fragmented nature of the contigs considerably affects its performance metrics (sensitivity: 38.77%, accuracy: 50.37%, F1 score: 30.16%, MCC: 25.65%). Lastly, PHACTS consistently underperforms, regardless of whether it is assessing complete genomes or simulated assembled contigs.

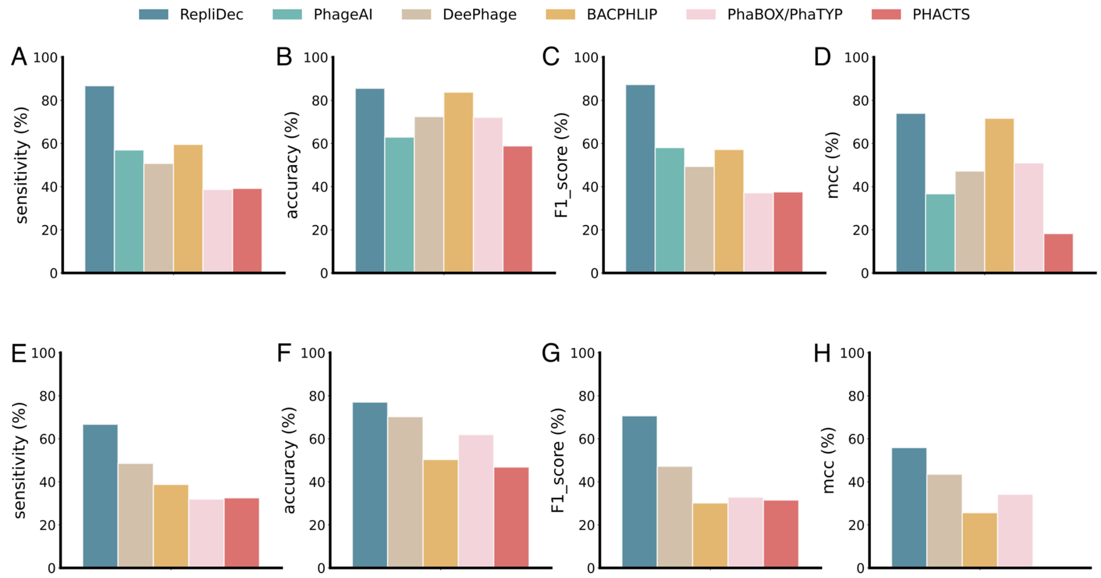


Figure 3-5 **Performance of different tools in predicting the lifecycle of bioinformatic predicted test dataset (Test Dataset 2).**

Comparison of complete phage genomes between different tools on sensitivity(A), accuracy(B), f1_score(C), and MCC (D). Comparison of assembled contigs simulated from test dataset 2 genomes between different tools on sensitivity(E), accuracy(F), f1_score(G), MCC (H).

3.1.6 Temperate phage abundance in patients with Inflammatory Bowel Disease (IBD) using RepliDec+

To enhance the reliability of the replication cycle for complex metagenomics assemblies, I created an integrated pipeline called RepliDec+, which includes BACPHLIP, DeePhage, PhaBOX/PhaTYP, and RepliDec. RepliDec operates effectively in most scenarios. I utilized BACPHLIP, DeePhage, and PhaBOX/PhaTYP to improve performance in more complex situations. To prevent prediction conflicts between these tools, I developed a scoring system that calculates the confidence score for each input sequence’s two replication cycles (virulent/temperate). The weights of each tool were obtained based on the evaluation results of these tools from test dataset 1 and test dataset 2. RepliDec directly predicted the chronic replication cycle.

Next, I use RepliDec+ to predict the abundance of phages with different lifecycles in patients with IBD and a healthy group. I adopt a free assembly strategy and reference the Gut Phage Database (GPD). Initially, I analyzed about 142k non-redundant viral genomes in the GPD. Of these, 88,851 (62.22%) contigs were identified as Temperate, while 52,718 (36.92%) were classified as Virulent and 1,240 (0.87%) were categorized as Chronic. In accordance with temperate, phages constitute at least 20% to 50 % of free phages in the human gut (Sausset et al., 2020).

To investigate the role of temperate phages in human health, I further assess the abundance of temperate phages in 115 inflammatory bowel disease (IBD) samples by mapping them to the GPD database. It is evident that a greater number of patient samples with Crohn's disease (CD) and ulcerative colitis (UC) exhibit significantly higher abundances of temperate phages compared to healthy samples (Figure 3-6 A). The average abundance of temperate phages in CD patients is 17.18%, while in UC patients, it is 34.98%, both exceeding the levels found in the healthy cohort (16.42%). No statistically significant differences were identified between patients and healthy controls, as UC and CD patients were under different disease statuses, such as flare and early inactive phases.

Some patients with UC and CD, regardless of their state being mild, moderate, or severe, demonstrate a significantly high abundance of temperate phages (Appendix Figure 3), suggesting a potential link between these phages and IBD conditions. Our analysis revealed that the average abundance of temperate phages for CD patients and UC patients in "Moderate," "Severe," and “Mild” states is 35.29% (n=21) and 29.55% (n=23) respectively, both significantly higher compared to healthy individuals (pCD= 0.00036 and pUC=0.0055). Most samples exhibited very low levels of chronic phages, except for one moderate CD patient, whose abundance was mainly influenced by a single contig in GPD (uvig_277057, 13.37%). These temperate phages may play a crucial role in the overall gut microbiota and human health.

Patients with Ulcerative Colitis (UC) and Crohn's Disease (CD), whether their condition is mild, moderate, or severe, show a notably high presence of temperate phages (Appendix Figure 3). This suggests a possible connection between these phages and Inflammatory Bowel Disease (IBD). The average abundance of temperate phages is 35.29% in CD patients (n=21) and 29.55% in UC patients (n=23) for mild, moderate, or severe levels, both of which are significantly higher than in healthy individuals (pCD= 0.00036 and pUC=0.0055). Most samples had very low levels of chronic phages, except one moderate CD patient whose levels were primarily influenced by a single contig in GPD (uvig_277057, 13.37%). These phages may play an essential role in gut microbiota and overall human health.

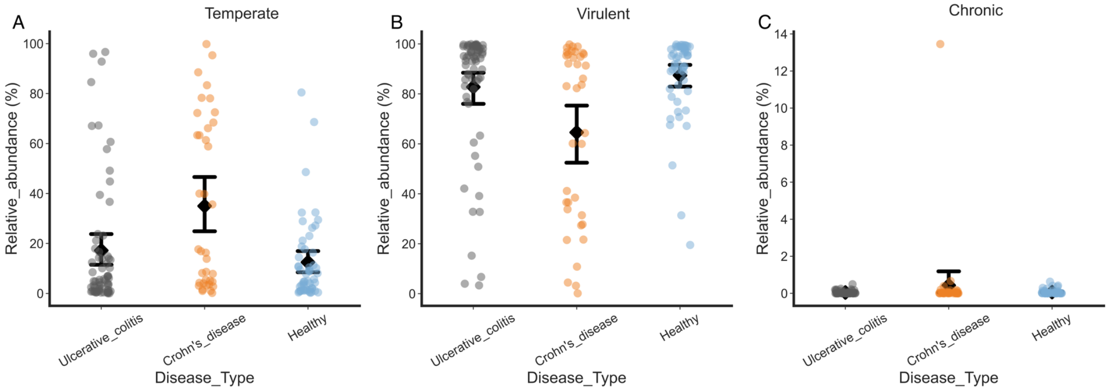


Figure 3-6 **Abundance profile in IBD and healthy cohort.**
Plots show the relative abundance of temperate phages (A), virulent phages (B), and chronic phages (C) in all healthy controls (household controls + healthy controls), CD, and UC samples. Error bars indicate the mean ± standard deviation (SD).

3.2 Temperate phage sequence detection in commercial cocktails

This chapter contains the results of a manuscript that has been submitted to bioRxiv preprint:

Xue Peng, Sophie Elizabeth Smith, Wanqi Huang, Jinlong Ru, Mohammadali Khan Mirzaei, Li Deng. 2024. Metagenomic analyses of single phages and phage cocktails show instances of contamination with temperate phages and bacterial DNA. bioRxiv 2024.09.12.612727; doi: https://doi.org/10.1101/2024.09.12.612727

Remarks: S.E.S. and W.H. isolated phages; S.E.S and X.P drafted the manuscript; X.P. performed the analyses; J.R. contributed to the analyses; M.K.M. and L.D. conceived and supervised the study and revised the manuscript. All authors reviewed and approved the manuscript.

3.2.1 Overview of commercial phage cocktails

Previously, the Eliava Institute sequenced four commercially available cocktails. Three of these are PYO cocktails (PYO97, PYO2000, and PYO 2014), which claim to target a broad spectrum of pathogens and are designed to treat infections related to burns, respiratory issues, gastrointestinal problems, and more. The fourth cocktail is the INTESTI, specifically formulated for gastrointestinal infections. The basic information about sequences is listed below.

Table 3-1 Relative abundance of temperate and virulent sequences in the four commercial cocktails.

	PYO97	PYO2000	PYO2014	INTESTI
Accession number	ERR2184199	ERR2184200	ERR2184201	SRR3744915
mean raw reads length	243	248	244	177
Total raw reads base	2 333 659 551	629 863 716	8 673 605 455	166 115 639
Total number of raw reads	9 575 268	2 533 708	35 454 908	938 802
Total clean reads base	2 057 343 861	249 631 025	7 458 026 778	157 824 205
Total number of clean reads	8 492 056	1 037 354	30 694 542	863 326
Total number of contigs (≥1kb)	470	247	755	225
Min length (bp)	1000	1005	1000	1003
Max length (bp)	344 583	93 033	212 402	128 699
Number of binned sequences	133	55	256	82
Number of bins	35	13	46	23
Total number of viral contigs	256	152	380	196
Number of viral contigs above medium quality	19	6	16	15
Average length of viral contigs above medium quality (bp)	8 5917.32	53 214.17	73 865.00	67 443.13
Maximum length of viral contigs above medium quality (bp)	344 583	93 033	212 402	128 699
Median length of viral contigs above medium quality (bp)	47 144	54 539	47 897	58 371
Minimum length of viral contigs above medium quality (bp)	5441	5485	38 562	28 300
Relative abundance of temperate viral contigs *	0.1674 (31)	0.0949 (4)	0.6998 (45)	1.8015 (9)
Relative abundance of temperate viral contigs with integrase gene *	0.0173 (2)	-	0.3926 (4)	-

*The absolute number of contigs appears in brackets following the relative abundance values.

I first investigate the taxonomy of all the assembled contigs (Figure 3-7 B), and PYO97 consisted of 15.21% *Herelleviridae*, 4.58% *Straboviridae*, 4.30% *Schitoviridae*, and 1.44% *Autographiviridae*. PYO2000 had 5.44% *Schitoviridae*, 1.51% *Autographiviridae*, 0.25% *Microviridae*, and 9.26% *Straboviridae*. PYO2014 was the most diverse cocktail, featuring 9.09%

Demerecviridae, 15.46% *Straboviridae*, 3.86% *Siphoviridae*, 5.64% *Autographiviridae*, and 3.03% *Herelleviridae*. In the INTESTI cocktail, 4.38% of the contigs were classified as *Autographiviridae*, 4.45% as *Straboviridae*, 6.38% as *Demerecviridae*, 10.56% as *Herelleviridae*, while 48.24% remained of unknown taxonomy.

Since most contigs cannot be assigned a taxonomy, I predict the host range for the cocktails (Figure 3-7 A). PYO cocktails are advertised to infect *Staphylococcus aureus*, *Streptococcus spp*. including *S. pyogenes*, *S.sanguis*, *S. salivarius* and *S. agalactiae*, *E. coli*, *Pseudomonas aeruginosa*, *Proteus mirabilis* and *Proteus vulgaris*. Metagenomic analysis of the phage cocktails showed that 39.56% of the assembled phages in the PYO97 cocktail were expected to infect *Staphylococcus* species. Additionally, 2.40% were predicted to infect *Pseudomonas* species, 0.27% *Escherichia* species, and 0.90% *Vibrio* species. For the remaining 41.07% of contigs, the host remains undetermined. In the PYO2000 cocktail, most contigs (12.26%) were predicted to target *Escherichia* species, while 6.10% were directed toward *Pseudomonas* species, and 37.64% were assigned to unknown hosts. Conversely, PYO2014 exhibited greater diversity, with 7.85% predicted to infect *Escherichia* species, 12.10% targeting *Enterococcus* species, 12.85% for *Enterobacter*, 0.23% affecting *Pseudomonas* species, and 9.09% infecting *Proteus* species. The host for the remaining 45.02% of contigs could not be identified.

INTESTI cocktail is designed to combat infections in the digestive system. It reportedly includes phages that specifically target *Shigella spp.*, *Salmonella spp.*, *E. coli*, *Proteus vulgaris*, *Proteus mirabilis*, *Staphylococcus aureus*, *Pseudomonas aeruginosa*, and *Enterococcus faecalis*. Notably, 53.44% of contigs could not be assigned a host. Infection predictions include 6.06% for *Escherichia spp.*, 2.57% for *Proteus spp.*, 1.46% for *Staphylococcus spp.*, and 4.42% for *Enterococcus spp.*.

3.2.2 Temperate phage sequences are present in commercial cocktails

The cocktails' metagenomic data were analyzed to check if they included temperate phages. (Table 3-1; Figure 3-7 C). RepliDec was employed to forecast the replication cycles of the viral contigs, resulting in 89 being categorized as temperate (Appendix Table 6). Additionally, I examined the presence of genes linked to a lysogenic lifecycle, such as integrases and excisionases. I detected 6 viral contigs containing an integrase gene – 2 in the PYO97 cocktail and 4 in the PYO2014 cocktail (Figure 3-7 C; Appendix Table 6). To attain absolute certainty, a comparative genomic analysis was undertaken to align each of these sequences with its nearest bacteria and phage sequences in the NT database identified through BLASTn, and amino acid comparisons were conducted using Clinker (Figure 3-7 C).

3.2.3 A low level of contamination was found in phage cocktails

I also investigated potential bacterial contamination in the phage cocktail sequences. They are mainly from *Escherichia coli* and *Shigella flexneri* in the PYO cocktail sequences. In contrast, the INTESTI cocktail showed minimal bacterial contamination, indicating differing contamination levels among the cocktails. The bacterial reads mapped across the complete bacterial genome, implying that the sequences are not exclusively from the prophage regions of the reference genomes but are likely a result of contaminating DNA rather than prophages. Nevertheless, the number of bacterial reads remaining after quality filtering in these cocktails is low.

3.2.4 Endolysin diversity from phage cocktails

I also investigate the diversity of endolysins present in viral contigs from different cocktails. In total, 74 endolysins were detected, and the endolysin phylogenetic tree clearly shows a high level of diversity among these genes. Two large endolysin clades and several small clades were observed. One significant clade identified in this study is the phage lysozyme (PF00959). Another large clade is Hydrolase 2, corresponding to PF07486. Some endolysins belong to a specific clade, showcasing remarkable similarities at the protein level. However, it is interesting to note that these endolysins originate from distinct cocktails, highlighting the diversity in their sources despite their shared characteristics.

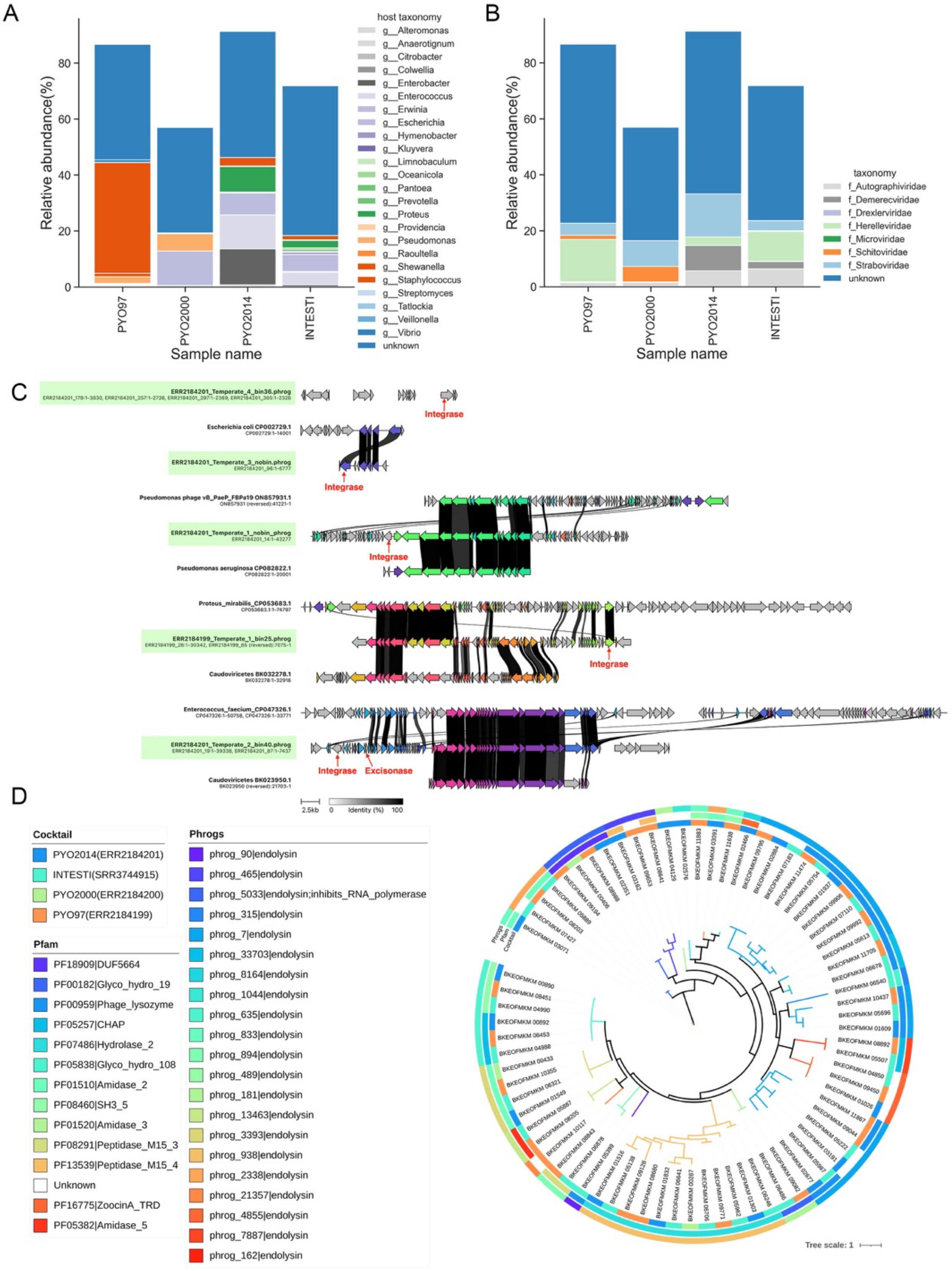


Figure 3-7 **Overview analysis of four phage cocktails.**

Information regarding predicted hosts (A) and taxonomy (B) was checked for four phage cocktails. (C) A comparative genomic map of four temperate sequences identified in cocktail sequences with the most similar sequences found in the NT database. The minimum identity of connections between genes is 0.6. Lysogeny-related genes are labeled. (D) endolysin tree built from sequences from phage cocktails.

3.3 Phage bacteria association detection using viral tagging in cross-over infection samples

This chapter contains the results of a manuscript that in preparation:

Xue Peng, Magdalena Unterer, Mohammadali Khan Mirzaei, Li Deng. 2025. Unveiling Bacteria and Phages Associations Using Viral Tagging on Cross-over infections samples. In preparation.

Remarks: Unterer M conduct all the wetlab work including collecting samples and viral tagging; Unterer M and X Peng drafted the manuscript; X.P. performed the bioinformatic analyses and visualize figures; M.K.M. and L.D. conceived and supervised the study and revised the manuscript.

3.3.1 Viral tagging in cross-over infection samples

Fecal samples from three healthy human subjects, three individuals diagnosed with ulcerative colitis (UC), and patients with colorectal cancer (CRC), including three in the early stages and two in the advanced stage, were collected. The bacterial and viral isolates obtained from each subject were pooled together according to their respective disease conditions, and subsequent cross-infection experiments were performed (Table 2-1). Utilizing the VT method established before (Unterer et al., 2023), 100 cells of each cross-over infection sample were sorted, and subsequent whole genome sequencing was conducted on these sorted cells. Since 100 cells were sorted and sequenced, they have highly uneven read coverage, which limits the assembly quality. To better understand the bacteria source, a non-redundant category of reference bacteria genomes was obtained from metagenome-assembled genomes (MAGs) (NRbacteria) assembled from native fecal metagenomes of each subject and a public bacteria database (Almeida et al., 2021, p. 204938). In total, 106,263 reference bacteria genomes from 841 bacterial genera were obtained. VT reads can map to 5,341 reference bacterial genomes from 208 bacterial genera, each with over 100 mapping reads and coverage greater than 10.

A similar strategy was applied to identify the viruses in VT. Because of the uneven coverage in VT reads, I decided to obtain the reference viral genome using multiple strategies. First, I conducted cross-assembly for virome, and VT reads to obtain the longer genomes. Then, an average of 19.31% of VT reads can map to the GPD database, so these GPD sequences can also be retained. Putative cross-assembly viral sequences derived from cross-assembly were retained if they contained at least one viral gene identified by CheckV. Virome and VT were also assembled separately to minimize the effects of different assembly strategies, and viral sequences were identified using VirSorter and CheckV. The single-assembly viral sequences were aligned to the cross-assembly sequences to serve as supplementary viral sources for the reference viral genome.

However, according to the new ICTV viral taxonomy classification system (Turner et al., 2023), all the viral genomes are identified as *Caudoviricetes* and *Malgrandaviricetes*. So, I cluster the viral sequences, including cross-assembly viral sequences, single-assembly viral sequences, and GPD sequences, into viral clusters (VCs). In total, 7336 VC at strain level (98% ANI) and 4395 VC at species level (95% ANI) were obtained. However, phages that show high genomic similarity can exhibit differences in host specificity. For example, seven Bacillus subtilis phages revealed high nucleotide and amino acid similarity. However, the host range is slightly different (Loney et al., 2023). So, I focus on strain levels (98%ANI). And VT reads can map to 607 strain-level VCs.

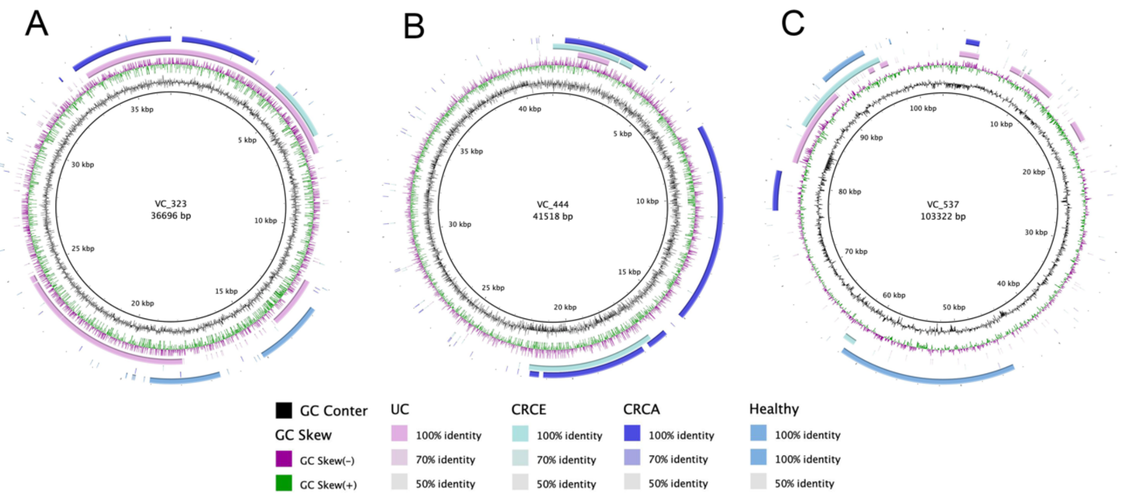


Figure 3-8 **Examples of viral clusters detected in multiple disease conditions.**
Each ring represents scaffolded phages assembled from different disease conditions, as indicated by the color code in the figure legend.

3.3.2 Bacteria profile in UC, CRC and healthy control

The overall relative abundance of identified bacterial taxa is lower in patients than in healthy controls (H) (Appendix Figure 4 B). Regarding bacterial composition, *Escherichia* represented 48.26% in healthy controls. However, its prevalence significantly decreased to 1.07% in ulcerative colitis (UC) patients. It was also observed at 3.28% in early-stage colorectal cancer (CRC) patients (CRCE) and at 6.47% in those with advanced-stage CRC (CRCA). Statistical analyses show a notable reduction of *Escherichia* in UC patients (p-value < 0.01) and in both early (CRCE: p-value < 0.01) and advanced stages (CRCA: p-value < 0.01) of CRC.

In healthy individuals, *Kluyvera* had a relative abundance of 3.19%; however, this was significantly reduced in patients with UC, CRCE, and CRCA. Conversely, *Bacteroides* showed a different trend, with a 6.76% abundance in healthy individuals, significantly lower than in UC (16.28%, p < 0.01), CRCE (21.77%, p < 0.01), and CRCA (29.61%, p < 0.01). A similar trend was noted for the genera *Prevotella* and *Phocaeicola*. As for *Bacillus*, the relative abundances were 4.17% in UC, 2.28% in CRCE, and 1.42% in CRCA, all significantly greater than the 0.04% found in healthy individuals (p < 0.01). Additionally, healthy individuals had a *Streptococcus* abundance of 1.54%, which was lower than the levels observed in UC (5.22%) and CRCE (2.10%), indicating that both UC and CRCE are associated with elevated levels of *Streptococcus* compared to healthy controls. In addition, *Elizabethkingia* is a ubiquitous pathogenic bacterium (Zajmi et al., 2022) that has been identified in CRC patients (C.-Y. Fang et al., 2021) and detected in our treatments, both at early stages (2.35%) and advanced stages (3.56%), compared to ulcerative colitis (UC) patients (0.35%) and healthy individuals (0.97%).

3.3.3 VT between different cross-over infection samples

Each cross-over infection VT sample exhibits significant variation in the dominant viral cluster (VC). In self-infection VT, the count of high-abundance VCs is lower than the cross-infection group (Appendix Figure 5 A). This may result from a broader diversity of bacterial sources (Appendix Figure 4). *Escherichia* is the primary bacterial taxon in self-infection samples (Appendix Figure 5 B). Furthermore, 29.32% (178 VCs) of the identified VCs in VT are predicted to be *Escherichia* phage, according to CRISPR and iPHoP predictions. VC 173 is one of the *Escherichia* phages that exhibit a high abundance within cross-infection samples. It is found in 52.50% of Hb-Hv, indicating a strong presence in healthy individuals. However, in samples from those with gut dysbiosis, VC 173 is much less common: it appears in only 10.97% of CRCEb-CRCEv samples and just 1.44% of CRCAb-CRCAv samples. A lack of proper bacterial host strains might cause the low presence of VC 173 in self-infection VT from CRC patients. When using the bacterial source from the healthy control instead of that from CRC to cross-infect the viral source from patients with CRC in the early and advanced stages, VC 173 exhibited high abundance in the early stages, comprising 80.72% of the sample (Hb-CRCEv). In the advanced stages, its abundance reached 94.04% (Hb-CRCAv). These findings indicate that VC 173 is abundant in the viral sources of colorectal cancer (CRC) patients and targets a specific group of *Escherichia* bacteria. The lack of specific bacterial hosts in the CRC bacterial source contributes to the low occurrence of VC173 within self-infection samples. In addition, a lower abundance of *Escherichia* in CRCEb-CRCEv and CRCAb-CRCAv was observed (Appendix Figure 5 B). One hypothesis indicates that VC 173 modulates the *Escherichia* population within the gut. In patients with CRC, the elevated levels of VC 173 appear to interact with the *Escherichia* host, leading to dysbiosis in the gut microbiome.

VC 146 is a potential phage targeting *Klebsiella*, primarily found in self-infections related to UC. It accounts for 83.16% within the UCb-UCv samples, significantly higher than in other self-infection samples, where it appears at only 0.28% in CRCEb-CRCEv (Appendix Figure 5 B). Conversely, VC 146 is rarely detected in cross-infection samples from either bacteria or vial sources of UC, such as Hb-UCv (0.71%), CRCEb-UCv (0%), CRCAb-UCv (0.04%), UCb-Hv (0%), UCb-CRCEv (0%), and UCb-CRCAv (0%). These observations indicate that VC 146 and its host are unique pairs linked to UC.

3.3.4 VT across various disease conditions

Although CRISPR and iPHoP can identify specific hosts of the VCs, their ability is limited. Using Netcomi to discover further phage-bacteria interactions in the cross-over infection samples, I create microbial association networks for each disease condition. Samples of cross-over infections, sourced from the same bacteria, were analyzed within one association network, shedding light on the dynamics of phage-bacteria shifts across varying disease conditions. In total, 607 VCs and 208 bacterial genera were used to construct this association network. The findings indicate that all cross-over infection samples contain 16 VCs and 18 bacterial genera in common (Figure 3-9). Notably, among the 16 VCs, 13 VCs (81.25% in total) are capable of infecting *Escherichia* by the predicted host. Additionally, VC 39 targets *Citrobacter*, VC 290 infects *Salmonella*, and VC 297 impacts *Raoultella*.

In the healthy group, a total of 29 bacterial genera and 104 VCs formed 1,807 associations, as illustrated in Figure 3-9. Within this group, 56 VCs and two bacterial genera, *Acinetobacter* and *Rhodoferax*, stand out distinctly. Among these, 14 VCs contain the sequences only from healthy controls. Furthermore, four hub VCs and three hub bacterial taxa were identified from the association network (Table 3-2).

VC 445 exhibits the highest eigenvector centrality among the healthy control group, demonstrating a value of 1. This variable is identified within the CRCEb and UCb groups, possessing eigenvector centrality values of 0.18 and 0.49, respectively. VC 445 shows a strong association with multiple bacterial hosts (adjacency value = 1), including *Acinetobacter*, *Bacillus*, *Bacteroides*, *Blautia*, *Clostridium*, *Collinsella*, *Deinococcus*, *Dorea*, *Enterobacter*, *Escherichia*, *Fusicatenibacter*, *Massilistercora*, *Muricomes*, *Parabacteroides*, *Pseudomonas*, *Rothia*, *Salmonella*, *Slackia*, and *Streptococcus*. The predicted host for VC 445 is *Escherichia*, which is consistent with the results obtained from the network analysis. A similar observation is made with VC 318, for which the predicted host is also *Escherichia*, alongside significant associations with *Bacteroides* (adjacency value = 0.83), *Blautia* (adjacency value = 0.82), *Enterobacter* (adjacency value = 0.83), *Escherichia* (adjacency value = 0.88), *Fusicatenibacter* (adjacency value = 1), *Massilistercora* (adjacency value = 1), *Rothia* (adjacency value = 0.90), and *Streptococcus* (adjacency value = 0.81), all exhibiting adjacency values above 0.8 in the Hb group. VC 39, recognized as one of the hub VCs, has the second-highest eigenvector centrality of 0.898 among all nodes within the healthy group. It demonstrates a strong association with *Escherichia*, *Fusicatenibacter*, *Massilistercora*, and *Salmonella*, each possessing an adjacency value of 1 (Table 3-2; Figure 3-9 A). However, the predicted host organism is *Citrobacter*.

The UCb network (Figure 3-9 B) comprises 55 bacterial genera and 136 VCs. Collectively, these bacteria and VCs establish a total of 4,214 associations between phage and bacteria pairs. Notably, three central bacterial strains exhibit extensive connections to the majority of the VC nodes (Table 3-2). *Massilistercora* demonstrates the highest eigenvector centrality score of 1, indicating a strong association with 79 VCs (adjacency value = 1). *Escherichia* follows closely, with an eigenvector centrality of 0.82, showcasing robust connections to 48 VCs (adjacency value = 1). Furthermore, *Monoglobus* is linked to 35 VCs with an adjacency value exceeding 0.8. Four bacterial taxa, specifically *Proteus*, *Aeromonas*, *Ezakiella*, and *Desulfotalea*, are uniquely identified within the UC group. Additionally, 66 VCs are distinct to the UC group, of which 16 VCs contain sequences assembled solely from the UC.

In contrast to the healthy cohort, CRCE (Figure 3-9 C) and CRCA (Figure 3-9 D) exhibited an increase in the number of bacteria and viral communities (VCs). This observation indicates a significant diversity of bacteria and phages among patients diagnosed with CRC. A total of 8,382 associations were identified within the CRCE group, comprising 87 distinct bacterial taxa and 142 VCs. Notably, 75 VCs and 25 bacterial taxa are considered unique among these. *Desulfitobacterium*, *Helicobacter*, and *Thermotoga* represent three key taxa that are exclusive to the CRCE. *Helicobacter* possesses an eigenvalue centrality of 0.80 and is associated with 124 VCs, of which 57 VCs demonstrate medium to strong association, as indicated by adjacency values exceeding 0.5. Furthermore, it is significant to note that more than half of the hosts for these VCs remain unidentified. *Fusicatenibacter*, recognized as one of the hub bacterial taxa, exhibits a strong association with multiple VCs (62 VCs), presenting an adjacency value of 1, and has not been detected in the UCb and CRCAb groups.

In the CRCA group (Figure 3-9 D), 6,212 associations are established by 81 bacterial genera and 131 VCs. Among these, 81 VCs and 29 bacterial genera are exclusive to the CRCA, while 19 VCs are solely derived from the CRCA. VC 1, VC 24, and VC 28 demonstrate high eigenvector centrality and only from CRCA without knowing their host. These three VCs demonstrate robust connections to *Acetivibrio*, Calothrix, *Collimonas*, *Faecalimonas*, *Fermentimonas*, *Mageeibacillus*, and *Roseburia*, with adjacency values exceeding 0.8.

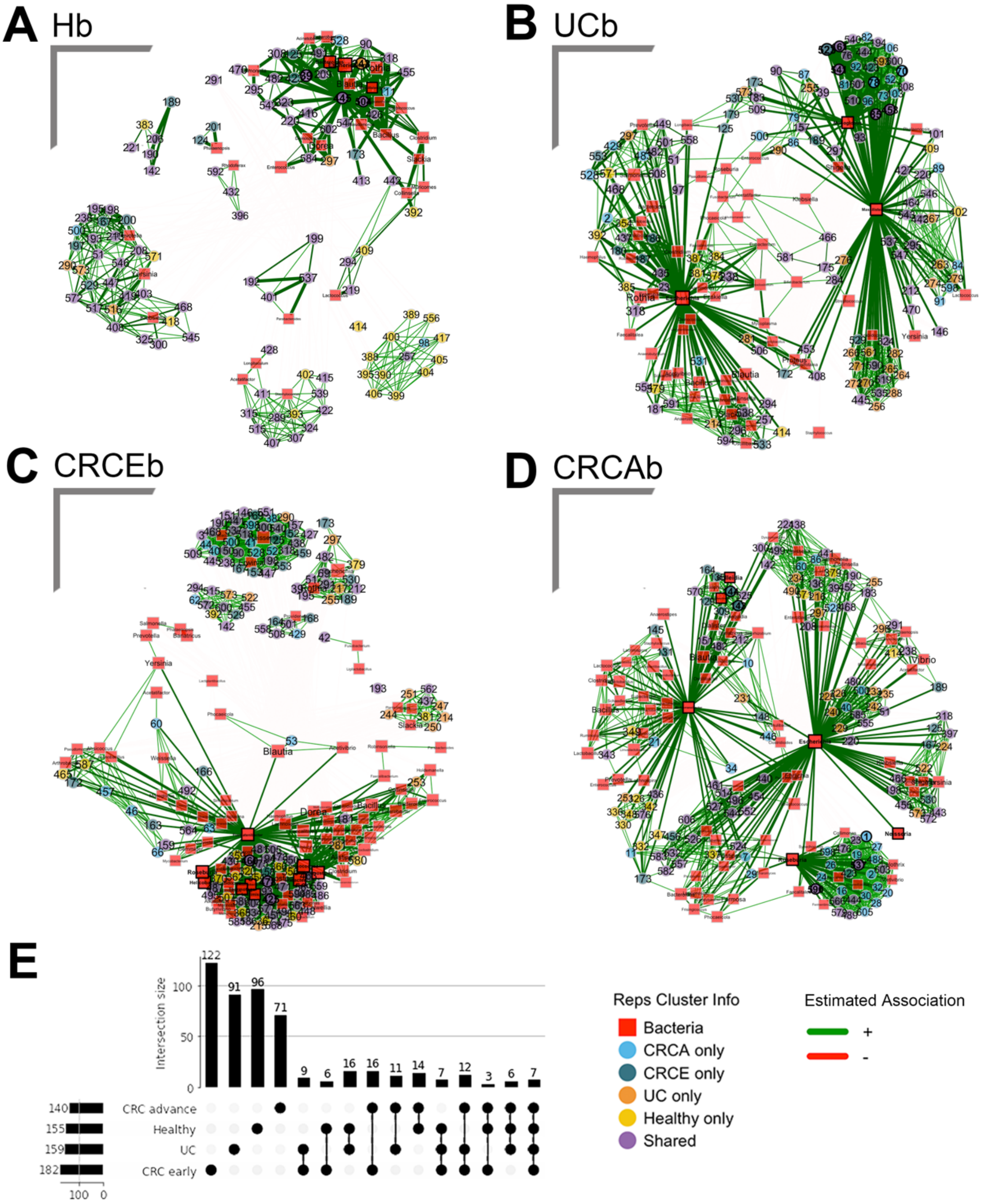


Figure 3-9 **VCs and bacteria differ in disease conditions.**

Phage-host association network in Hb(A), UCb(B), CRCEb(C), and CRCAb(D).

Green edges indicate positive estimated associations, while red edges signify negative ones. Eigenvector centrality determines hubs (nodes with a centrality value exceeding the empirical 95% quantile) and scales node sizes. Hubs are emphasized with bold text and borders. Node colors represent different categories. (E) The UpSet diagram displays how the VCs vary across disease conditions.

Table 3-2 Hub nodes have been identified within each disease group

	Hub VC	Hub bacteria taxa
Hb	VC 39, VC 243, VC 445, VC 504	Bacteroides, Escherichia Massilistercora
UCb	VC 70, VC 78, VC 85, VC 161, VC 458, VC 523, VC 541	Escherichia, Massilistercora, Monoglobus
CRCEb	VC 425, VC 460, VC 478	Anaerostipes, Deinococcus, Desulfitobacterium, Eubacterium, Fusicatenibacter, Helicobacter, Lacrimispora, Roseburia, Thermotoga
CRCAb	VC 1, VC 143, VC 144, VC 537, VC 590	Bulleidia, Erysipelatoclostridium, Escherichia, Mediterraneibacter, Neisseria, Roseburia

4. Discussion

Generally, phages are thought to undergo one of three replication cycles: lytic, lysogenic, or chronic. Nonetheless, the specifics of phage replication mechanisms remain unclear. This highlights the intricacy of microbial interactions and the need for innovative research methods to uncover these crucial yet elusive biological processes. Currently, experimental methods remain the gold standard for determining the replication cycle (Edwards et al., 2016; Versoza & Pfeifer, 2022). Testing every phage strain is impractical due to the time and effort required, coupled with the fact that not all phages can thrive in lab environments. While many computational tools exist, they predominantly predict only lytic and lysogenic cycles. This limitation can result in the misclassification of chronic phages, particularly when examining their lifecycle within mixed communities, such as those from metagenomes or viromes. To reduce the misclassification of chronic phages in phage-host relationship studies, I developed RepliDec, which accurately predicts lytic, lysogenic, and chronic cycles.

Temperate phages, or prophages, represent the direct interactions between phages and bacteria. Nevertheless, a comprehensive understanding of the interaction mechanisms of temperate phages remains elusive. In this context, I employ RepliDec+ and RepliDec to investigate the role of temperate phages in real-world scenarios, particularly focusing on the gut microbiome and commercially available phage cocktails. Within the gut environment, phages, bacteria, archaea, and fungi coexist, potentially achieving a state of homeostasis. I assess the relative abundance of temperate phages in patients diagnosed with gastrointestinal disorders in comparison to healthy control subjects. Utilizing RepliDec and RepliDec+, I can more accurately evaluate the impact of temperate phages on gut disorders.

Furthermore, I examine the influence of temperate phages on phage cocktails, which is one of the most successful applications of phages. Because prophages can facilitate the transfer of genetic material, potentially introducing deleterious elements in patients. Additionally, prophages may affect the efficacy of phage cocktails and complicate the determination of their concentrations during treatments. I investigate the presence of temperate phages and bacterial contamination in commercially available cocktails, which could provide essential insights for manufacturers and researchers, highlighting the need for increased scrutiny regarding temperate phages and their implications for the effectiveness and safety of phage therapy applications. This emphasizes the necessity for rigorous screening and monitoring of temperate phages within bacteriophage therapy and its modification.

One significant challenge associated with studying PHI is the identification of specific phage-host pairs. Accurately delineating these pairs is crucial for enhancing our comprehension of phage dynamics, host specificity, and the overall impact of phages on microbial communities. Although bacteriophages typically exhibit a relatively narrow host range, identifying all host bacterial strains poses a considerable challenge. Currently, experimental methodology is still the gold standard for recognizing PHI, with plaque assays being the most prevalent technique employed (Edwards et al., 2016; Nie et al., 2024). However, using bacterial strains in plaque assays is difficult, as it is impracticable to test every bacterial strain due to the considerable time and workload. Moreover, it is essential to note that not all bacterial strains can be cultured under laboratory conditions. Although various computational approaches exist, none have demonstrated the capacity to predict the comprehensive host spectrum accurately (Nie et al., 2024; Roux et al., 2022; Villarroel et al., 2016). Typically, predictions are limited to the genus or species level; however, phages infect specific bacterial strains in practice, and their efficacy can vary among different strains. For instance, the T4 bacteriophage is a well-documented lytic phage that infects *Escherichia coli.* Yet, not all *E. coli* strains exhibit susceptibility to T4, including *E. coli DFB1655 L9* (Lee et al., 2018; Stanton et al., 2023).

Utilizing an experimental approach employing viral tagging to identify phage-host pairs in patients with gastrointestinal disorders, comparing these findings to those from healthy controls. The adoption of viral tagging enabled me to circumvent the labor-intensive culturing processes typically required in laboratory settings, thereby preventing the loss of uncultured phages and bacteria. This modification (Unterer et al., 2023) significantly improves the capacity to capture phage-host pairs, thereby advancing our understanding of phage-bacteria interactions at a community scale. Concurrently, I developed a corresponding bioinformatics analysis workflow to identify phage-host pairs comprehensively. I utilized metagenomic and virome data as supplementary resources to construct an extensive database of bacterial and phage sources. By relying on the physical attachment of phages and bacteria captured through viral tagging techniques, I could establish the phage-host association network, elucidating the associations between specific phage-host pairs. This method not only enhances our understanding of phage-bacteria dynamics but also makes a substantial contribution to the fields of microbial ecology and phage therapy research.

This study comprehensively examines the interaction between phages and bacteria, with particular emphasis on temperate phages and prophages. I incorporated the chronic lifecycle into lifecycle predictions and assessed the impact of temperate phages in both practical applications and clinical samples. Furthermore, applied with a modified version of the viral tagging technique, alongside the bioinformatics analysis framework, enables the large-scale identification of phage-bacteria pairs without the need for culture. Such advancements may hold significant implications for the field of microbiology, particularly in the development of bacteriophage-based therapeutic applications.

4.1 RepliDec and RepliDec+

4.1.1 VT across various disease conditions

Bacteria and archaea can be quantified through 16S ribosomal RNA analysis in unknown communities (J. S. Johnson et al., 2019); however, quantifying bacteriophages is more challenging due to the absence of definitive hallmark genes (Hatfull, 2008). Moreover, understanding qualitative temperate phages or prophages remains a critical scientific question needing resolution. Currently, many software applications, such as VirSorter (Guo et al., 2021; Roux et al., 2015) and VIBRANT (Kieft et al., 2020), can predict the prophage regions in bacterial genomes. Still, fewer tools are available for predicting phage lifecycles, as this requires more complexity than prophage prediction. The genomic content between prophages and bacteria varies, including differences in GC skew (Grigoriev, 1998). Additionally, phages exhibit a higher gene density than bacteria (McNair et al., 2019), and prophages contain structural genes like capsids and tail genes, which are helpful in identifying their presence within bacterial chromosomes. In contrast to prophage prediction, there are fewer signals for lifecycle prediction.

Integrase and excisionase are critical biomarkers of lysogenic phages, facilitating the integration and excision of prophages from bacterial genomes. The integrase enzyme identifies the integration site within the bacterial genome by recombining the *attP* (phage) and *attB* (bacteria) attachment sites, resulting in the formation of *attL* (left) and *attR* (right) sites (H. Li et al., 2018).

A phage is typically classified as a temperate phage when two specific biomarkers are identified in its sequences. While this methodology can effectively analyze the complete phage genome, it may lack specificity in distinguishing sequences derived from virome data. For instance, it is important to note that the integrase is not exclusively encoded by temperate phages; integrative plasmids, pathogenicity islands, and both conjugative and mobilizable elements can also encode this protein (Williams, 2002). Additionally, due to the mosaic feature of the phage genome, it is challenging to assemble them to complete the genome level from a complex environment sample. Instead, they are fragmented pieces from the complete phage genomes (Roux, Adriaenssens, et al., 2019; Smits et al., 2014). For example, CrAss-like phage (crassviruses) is discovered by cross-assembling reads in human fecal metagenomes (Dutilh et al., 2014). These fragmented sequences probably do not contain two specific biomarkers, potentially leading to incorrectly classifying a temperate phage sequence as virulent. Consequently, none of the existing lifecycle prediction tools depend only on these lysogenic phage biomarkers.

RepliDec not only contains integrase and excisionase but also includes specific genes that have been utilized as biomarkers in other studies, including transposase, resolvase, and recombinase (Muscatt et al., 2022; Tang et al., 2023). In Mavrich's study (Mavrich & Hatfull, 2017), researchers employed "ParA", a gene belonging to the ParABS system, as marker genes for identifying temperate phages. The ParABS system, which is found in the Mycobacterium smegmatis bacteriophage RedRock, facilitates the formation of stable lysogens that harbor extrachromosomal prophages. This stability is contingent upon the Type Ib partitioning system, which encodes the proteins ParA and ParB, thereby ensuring the stability of the prophage (Dedrick et al., 2016). RepliDec broadens its scope to include proteins associated with the ParABS system and proteins that play a role in the CI-CII-Cro system. The CI-CII-Cro system is one of the most thoroughly characterized lytic-lysogenic switch systems, in which the CI and Cro regulators determine the lysogenic and lytic states, respectively, functioning as a bistable genetic switch (Oppenheim et al., 2005; Shao et al., 2019). CII, a lysogeny-promoting protein, inhibits the lytic state upon activation and establishes the lysogenic state by activating transcription from three specific promoters (pI promoter, pRE promoter, paQ promoter) (Murchland et al., 2014; Shao et al., 2017). The pI promoter facilitates the expression of integrase, while the pRE promoter activates the transcription of the repressor CI. During the transition from lysogenic to lytic phases, Cro indirectly reduces CII levels, thereby activating CI transcription (Schubert et al., 2007).

Using the PC tendency toward temperate, RepliDec also found some proteins indirectly related to the lysogenic process, such as DNA methyltransferase and UvsX-like recombinase. Methyltransferase takes part in the lytic process with CI protein and is essential for maintaining prophage lysogeny. The gene for the DNA methyltransferase (DNA MTase) located in the lysogenic module (the gp27 gene) encoded an enzyme with homology to MTases that modify the N6 position of adenine (N6-methyladenine [m6A]) (Smith & Jeddeloh, 2005). The adenine methyltransferase (DAM) methylates the rha antirepressor gene, and once methylation is removed, homologous CI repressor protein becomes repressed and non-functional, leading to the switching to the lytic cycle (Bochow et al., 2012; Murphy et al., 2008; Smith & Jeddeloh, 2005; Song, 2020; Yuan et al., 2015). In addition, UvsX-like recombinase is a 44 kDa protein that forms helical assemblies on to single-stranded DNA (ssDNA) and mediates DNA strand exchange between homologous chromosomes (Farb & Morrical, 2009). From previous studies, its crystal structure and the overall architecture and folds closely resemble that of RecA (Gajewski et al., 2011), which indicates that UvsX and RecA have similar functions. In SOS response, RecA binds to single ssDNA at the site of the lesion, forming active RecA-ssDNA complexes (RecA*) that promote autoproteolytic cleavage of prophage repressors as well as of the SOS repressor LexA (Rozanov et al., 1998). Similarly, UvsX-like recombinase can also form helical assemblies on ssRNA, which might also promote the prophage induction.

4.1.2 Comparison between lifecycle prediction tools

While seven tools exist to predict phage lifecycles, a systematic comparison among them is lacking. This gap arises from the limited experimental validation of phages' replication cycles. Currently, a phage dataset created by Mavrich et al. (Mavrich & Hatfull, 2017) and the Actinobacteriophages database (https://phagesdb.org/) (Russell & Hatfull, 2017) contains the experimentally validated lifecycle. However, the Actinobacteriophages database contains phages that only infect *Actinobacterial* hosts, leading to severe database bias when used as a training dataset. So, only the phage dataset created by Mavrich et al. (Mavrich & Hatfull, 2017) can be used as a training dataset. However, PhagePred, BAPHLIP, DeePhage, and PhaBox/PhaTYP utilize this dataset as a training dataset to train the machine learning model to predict the replication cycle. The lack of an experimental validated testing benchmark dataset makes it difficult to evaluate the performance of each tool fairly.

Additionally, the original articles on these tools contain small-scale comparisons. However, the varied test datasets and evaluation metrics chosen by the authors complicate the comparison of evaluation results across the papers. To resolve this challenge, I propose a new evaluation matrix named the “2-2-4” framework, encompassing two datasets, two data conditions, and four measurements. Test dataset 1, collected by Mavrich et al. (Mavrich & Hatfull, 2017) , contains 470 RefSeq phage genomes with experimentally validated replication cycles consistent with earlier in silico predictions. It contains 207 temperate, 261 virulent, and two chronic viruses (Appendix Table 3). To prevent overfitting issues and ensure a fair comparison of replication prediction tools, an extra-standard test dataset (Test dataset 2) has been created. Test dataset 2 includes 610 representative NCBI phage genomes with no overlap with RefSeq and dataset 1. The maximum similarity within test dataset 2 is 95%. It comprises 328 temperate, 253 virulent, and 29 chronic genomes (Appendix Table 3). The two data conditions show that I assessed these tools at both the complete phage genome level and the fragmented contig level, utilizing simulated metagenomics data sourced from the two datasets individually. Due to the variability of chronic phages in the testing dataset relative to temperate and virulent phages, I employ four metrics: sensitivity (Sn), accuracy (Acc), and two comprehensive metrics—F1-score and the Matthews correlation coefficient (MCC). These metrics collectively offer a thorough performance assessment, ensuring reliable evaluation results.

According to the evaluation results from the “2-2-4” framework, the overall performance in simulated contigs is worse than that in complete genomes, regardless of whether it is test dataset 1 or test dataset 2. This indicates that the length of sequences could influence the prediction accuracy. Due to the mosaic feature of phages, the assembled contigs typically range from short fragments to nearly complete genomes (Roux, Adriaenssens, et al., 2019; Smits et al., 2014). As the contig length decreases, the available informative data for prediction may result in misclassification. Additionally, contigs from the same phages can have different prediction outcomes. At present, none of these tools effectively predict the replication cycle of short phage contigs.

Nonetheless, employing binning techniques presents a promising approach, enabling the aggregation of short contigs into high-quality phage bins. For example, researchers recovered 6,077 high-quality genomes from 1,024 viral populations via binning (Johansen et al., 2022). It is uncommon to bin phage contigs into viral metagenome-assembled genomes (vMAGs), instead choosing viral operational taxonomic units (vOTUs) derived from clustering phage sequences. This is primarily due to the many bins that may encompass fragmented phage contigs, which prove challenging to connect as one using overlapping regions. Furthermore, these bins may also include other mobile genetic elements, such as plasmids, potentially leading to erroneous conclusions in subsequent analyses (Joachim, 2022; Maguire et al., 2020). To avoid the potential drawback of binning, post-processing is a crucial step after binning to remove ambiguous viral bins and minimize the impact of false positive phage sequences (Joachim, 2022). Binning outputs can be directly utilized as inputs in RepliDec. In RepliDec+, additional scripts have been created to ease the use of binning outputs across all included software.

All tools performed worse on test dataset 2 compared to test dataset 1. The first dataset comes from RefSeq. These tools are likely trained on partial or complete RefSeq genomes, leading to potential overfitting. Improved performance is observed when phages closely resembling the genome are included in the training database. However, genomes in test dataset 2 show at most 95% similarity with genomes in RefSeq. This suggests that they differ from the training dataset to some extent. They are considered "novel" phages compared to the genomes in the training data, which can avoid the potential issues in test dataset 1. I observed a significant drop in the four metrics compared to test dataset 1, particularly in the simulated contig from test dataset 2. One explanation is that all these tools, including RepliDec, lack robustness because of the limitation of the training dataset. Approximately 3,000 RefSeq genomes at most have been used as the training dataset in available tools, which is far less than the estimated 10^31 viruses in the world (Comeau et al., 2008; “Microbiology by Numbers,” 2011). And only about 500 genomes have been experimentally validated. This is a very small part of the entire phage population in the world. The limited diversity in the training dataset leads to poor prediction performance against new phages. Expanding this diversity is a complex task, as collecting and sequencing phages from environmental sources can take years. Additionally, considerable effort is required to isolate phages from their natural environments and their bacterial hosts, which must then undergo experimental validation for their replication cycles. These steps can require significant time and financial investment. RepliDec uses a different strategy to increase the diversity in the training dataset. It not only contains about 4000 RefSeq genomes but also 21,134 prophages sequences. Prophage sequences can significantly increase the diversity of the training dataset, as prophage recombination enables the coexistence of multiple phage types (Nadeem & Wahl, 2017). So, RepliDec can still maintain the best performance among all the tools in all four metrics. Additionally, it is worth investing more effort in researching PHI, especially the lytic-lysogenic interaction. This focus may facilitate the identification of additional biomarkers or genes related to lysogenic systems, thereby enhancing the accuracy of predicting "novel" phages.

RepliDec also has limitations because it does not take into account the co-occurrence relationships of proteins in the lytic-lysogenic interactions. Naïve Bayes classifier assumes that each protein is independent. However, numerous biological processes are regulated by the collective actions of multiple genes (Erez et al., 2017; Hatfull & Hendrix, 2011; Ofir & Sorek, 2018). For example, the proteins CI, CII, and Cro collaboratively control lysis-to-lysogeny decisions in phage λ (Oppenheim et al., 2005). Regrettably, RepliDec does not integrate this association into its predictive model, primarily due to the difficulty of quantifying the interactions of these proteins into mathematical models. Moreover, there is still much "dark matter" in the multi-gene interaction systems.

Despite RepliDec using the largest training dataset of all available tools, many proteins still lack a thorough understanding of their functions. This gap in knowledge poses difficulties in correctly linking these proteins to specific biological processes, ultimately obstructing progress in the field.

4.1.3 Temperate phages in IBD patients

Phages are thought to play a vital role in both natural ecosystems and human health. They interact with bacterial hosts, regulate the richness and variety of bacterial populations, and affect bacterial metabolic processes. Phages primarily replicate through three cycles: lytic, lysogenic, and chronic (Hobbs & Abedon, 2016). Distinguishing between temperate and virulent phages is difficult due to the absence of marker genes, especially in viral contigs constructed from viromic and metagenomic data. Existing in silico tools struggle to accurately predict the replication cycles of chronic phage sequences, potentially leading to faulty predictions. Furthermore, the limited diversity in the training datasets used by these tools hampers their ability to reliably identify temperate phages that share only remote homology with the dataset.

To address these issues, I develop RepliDec and an integrated pipeline. RepliDec+. Then, I re-evaluated the temperate phage in patients with UC, CD, and healthy controls. The gut is a complex ecosystem comprising bacteriophages, bacteria, and fungi. Investigating temperate phages in the gastrointestinal tract is promising, as around 80% of intestinal bacteria contain prophages. Additionally, studies suggest that temperate phages account for 20% to 50% of the free phages in the gut (Henrot & Petit, 2022; Sausset et al., 2020). From my study, I found that the average abundance of temperate phage in CD patients and UC patients in the "Moderate," "Severe," and “Mild” states was significantly higher than that in the healthy cohort.

This aligns with findings from other studies. Research indicates that the relative abundance of temperate phages varies between patients with IBD and those without (Nishiyama et al., 2020). Specifically, the temperate bacteriophages that infect *Bacteroides uniformis* and *Bacteroides thetaiotaomicron* were found to be overrepresented in patients with active UC (Nishiyama et al., 2020). Furthermore, Clooney et al. (Clooney et al., 2019) also reported an increased relative abundance of temperate phages in patients with UC and CD compared to healthy controls. Additionally, some temperate phages targeting *Faecalibacterium prausnitzii* exhibit higher occurrence and proportion in IBD patients than in healthy controls (Y. Cao et al., 2014).

Temperate bacteriophages constitute a significant portion of the human gut microbiota and are likely involved in the pathogenesis of IBD. Temperate phages might kill beneficial bacteria strains in the gut, leading to dysbiosis of the entire microorganism population. For example, *Bacteroides uniformis* and *Bacteroides thetaiotaomicron* are two bacteria species that have been experimentally proven to be beneficial to gut homeostasis (Nishiyama et al., 2020). Researchers discovered that in patients with IBD, there was a greater relative abundance of phages targeting these two bacteria, while their hosts were under-represented (Nishiyama et al., 2020). Also, *Faecalibacterium prausnitzii* are generally less abundant in IBD patients than in healthy controls. There is an observed increase in the relative abundance of *Faecalibacterium prausnitzii* phage in IBD patients, indicating a potential correlation between this phage activity and the depletion of *Faecalibacterium prausnitzii* (Y. Cao et al., 2014). In addition, IBD patients often experience an inflammatory environment in their gut, and the inflammatory environment can increase prophage induction (Clooney et al., 2019; Diard et al., 2017). When prophage induction occurs, temperate phages shift into the lytic cycle, leading to the lysis of beneficial bacteria. This exacerbates dysbiosis and perpetuates the inflammatory cycle. These findings indicate that bacteriophages significantly contribute to the pathogenesis of IBD by attacking and removing beneficial bacteria.

4.2 Phage cocktails and safety

The global surge in antibiotic-resistant (AMR) bacterial strains has renewed interest in phage therapy as an alternative treatment. For effective therapeutic use, phages must be strictly virulent, efficiently targeting the bacterial host, and ideally, they should be fully characterized without any virulent genes (Strathdee et al., 2023). Utilizing newly developed tools like RepliDec, I re-examined the temperate phage fragments in four commercial cocktails. My analysis revealed that the PYO cocktail contains some temperate phage fragments, although their relative abundance is low. To date, no studies have evaluated the effects of temperate phages in these commercial cocktails, particularly concerning potential undesirable or harmful impacts on recipients. Additionally, the concentration at which these temperate phages might cause adverse effects remains unknown.

It is risky to apply the temperate phage in the cocktail because they often carry some virulent genes and AMG, which may promote the rapid evolution of bacterial pathogens, aiding their adaptation to bacteria host environments and clinical treatments (Davies et al., 2016). The temperate phage Pf, which infects *Pseudomonas*, can significantly enhance the virulence of its host infections (Secor et al., 2020). In addition, prophages lacking virulence genes can still have increased virulence on host bacteria, such as bacteriophage vB_Saus_PHB21. This phage was isolated from epidermal samples of Siberian tigers (*Panthera tigris altaica*) using a strain of *Staphylococcus aureus* (MRSA) known as SA14. Bacteriophage vB_Saus_PHB21 was a temperate bacteriophage belonging to the *Siphoviridae* family, and it did not contain any virulence genes. Nonetheless, integrating the PHB21 genome into the MRSA host significantly enhances the bacterium's cell adhesion capabilities, phagocytosis resistance, and biofilm formation. Consequently, this integration increased mortality rates in both *Galleria mellonella* and mouse models (D. Yang et al., 2022).

The use of temperate phages in phage therapy is not strictly prohibited. A phage cocktail consisting of four temperate phages, including PHB22a, PHB25a, PHB38a, and PHB40a, isolated from environmental sewage, has been used to instigate biofilms and treat Methicillin-resistant *Staphylococcus aureus (MRSA) S-18* infection in animal models. Treatment with this phage cocktail containing Ca^{2+} and Zn^{2+} improved the survival of *Galleria mellonell* larvae against Methicillin-resistant *Staphylococcus aureus (MRSA)* S-18 infection. It also reduced the bacterial load of tissues in mice models (X. Li et al., 2022). Furthermore, phage engineering techniques can be applied to temperate phages, enabling their use as lytic phages in phage cocktails. Using Bacteriophage Recombineering of Electroporated DNA (BRED) allows for modifying therapeutic phages from a temperate to a lytic state. In a case study of a cystic fibrosis patient, two engineered temperate phages and one lytic phage were administered, resulting in observable signs of recovery after six months of treatment (Payaslian et al., 2021).

4.3 Phage-host associations with gut intestinal disease

Identifying the page-host pair is critical during PHI studies, while the experimental method, plaque assays, remains the gold standard for recognizing PHI (Edwards et al., 2016; Nie et al., 2024). In addition, the gut interaction is a multi-layer complex network involving the mammalian host, bacteria, virus, and fungi. It is challenging to investigate PHI in the gut environment via experimental methods. Here, relying on a modified version of viral tagging (Unterer et al., 2023) and whole genomic sequencing, I can obtain the associations between phages and bacteria captured through cross-infection samples using a modified version of viral tagging and whole genomic sequencing.

Numerous studies have confirmed a shift in the gut microorganisms community of patients compared to that of healthy individuals (Clooney et al., 2019; Duerkop et al., 2018; Loh & Blaut, 2012; Zheng et al., 2024). However, establishing the phage bacteria pairs from the entire community is challenging. VT methods have been used to identify the 363 unique host–phage pairings from the fecal microbiome (Džunková et al., 2019). Using the VT method on cross-infection samples, I found 607 VCs and 208 bacterial taxa and they were used to uncover phage-bacteria associations across various disease conditions. Numerous associations were identified: 1,807 in healthy individuals, 4,214 in UC, 8,382 in early-stage colorectal cancer (CRCE), and 6,212 in advanced-stage CRC (CRCA). Notably, the healthy group exhibited fewer phage-bacteria associations than those with UC and CRC. This indicates that patients tend to have more intricate phage-host relationships, potentially tied to gut bacterial imbalances, a condition referred to as dysbiosis.

According to recent studies (Basha et al., 2022; Clooney et al., 2019; Federici et al., 2022; Kennedy et al., 2024; Mills et al., 2022), UC is linked to an imbalance in the gut microbiota and gut bacteriophages. Compared to the healthy control, bacteria such as *Acinetobacter, Klebsiella, Escherichia, Veillonella, Desulfovibrio,* and *Bacteroides;* bacteriophages such as *Escherichia phage and Enterobacteria phage and some temperate phages are enriched in UC patients* (Amos et al., 2021; Basha et al., 2022; Clooney et al., 2019; Federici et al., 2022; Kennedy et al., 2024; Mills et al., 2022). Meanwhile, UC patients show a lower abundance of *Faecalibacterium* and *Microviridae* (Al-Bayati et al., 2023; Y. Cao et al., 2014; Z. Cao et al., 2022; Machiels et al., 2014). *Bifidobacterium* was also detected in UC patients, while their changes are not the same all the time. Some studies have reported a decreased trend of *Bifidobacterium* compared to the healthy control (De Caro et al., 2016; Duranti et al., 2016; Kennedy et al., 2024), while others have shown the opposite results (D.-Y. Kang et al., 2023; Wang et al., 2014). Most of the bacteria afomentioned were detected in the association network.

Additionally, the association network identifies four bacterial taxa: *Proteus*, *Aeromonas*, *Ezakiella*, and *Desulfotalea*, which are unique to the UC group. *Proteus* are low-abundance commensals in the human gut with significant pathogenic potential (Hamilton et al., 2018). They are linked to a significant reduction in absorption processes at the small intestinal mucosal membrane, potentially worsening inflammation (Kanareykina et al., 1987). *Aeromonas* are Gram-negative rods known to cause a range of diseases. They can potentially trigger the development of de novo chronic colitis in patients without a prior history of IBD. Some case studies show that *Aeromonas* infections can mimic IBD, particularly UC when they lead to colitis (Ahishali et al., 2007). In a specific group of IBD patients, seven had been diagnosed with IBD before the onset of the *Aeromonas* infection, while five received their IBD diagnosis after the *Aeromonas* was identified (Pereira Guedes et al., 2023). Moreover, the abundance of *Desulfovibrio* was increased in cases of acute colitis compared to healthy controls (Rowan et al., 2010). These unique bacterial taxa may play a significant role in UC patients, and investigating the bacteria and viral communities associated with it might offer new insights for using phage therapy as a treatment option.

Patients with a long-standing history of UC and CD will have an increased risk of developing CRC (Sato et al., 2023). 104 VCs and 106 bacterial taxa were identified as unique to CRC, appearing in both early and advanced stages. Specific bacterial taxa linked to CRC include members from the phyla *Firmicutes*, *Bacteroidetes*, *Actinobacteria*, and *Proteobacteria* (Rebersek, 2021), as well as genera such as *Fusobacterium*, *Escherichia*, *Bifidobacterium*, *Clostridium*, *Roseburia*, and

Faecalibacterium (Chattopadhyay et al., 2021; Rebersek, 2021). *Roseburia* is identified as hub bacterial taxa in CRCE and CRCA. Alterations in the gut virome are evident in CRC patients compared to healthy controls, with significant reductions in *Enterobacteria* phages and crAssphages (Gao et al., 2021). Additionally, 22 viral genera, including *Inovirus*, *Orthobunyavirus*, *Tunalikevirus*, *Phikzlikevirus*, and *Betabaculovirus*, were found to distinguish CRC patients from controls (Nakatsu et al., 2018). Associations among these bacterial taxa were also identified within the association network, further highlighted through insights gained from VT methods.

Using VT methods allows for identifying more phage-host associations within a community without requiring prior knowledge of its composition; however, this approach has certain limitations. First, VT can only capture phages that attach to bacterial surfaces. Phages with similar attachment structures may adhere to bacteria but might not necessarily infect them, making it premature to conclude that identified bacteria-phage pairs truly interact. Additionally, multiple associations were observed for certain VCs, raising the question of whether each VC infects all candidate bacteria or just one. VT methods alone cannot resolve this uncertainty, indicating the need for further analyses, such as single-cell and transcriptome sequencing. Nonetheless, the advantages of VT methods outweigh their limitations.

4.4 Beyong this work

The topic of phage-host interactions is extensive; however, this study only focuses on the phage replication cycles and identifies phage-bacteria association pairs from virome data using the viral tagging (VT) method. Utilizing RepliDec and RepliDec+, I evaluate the role of temperate phages in IBD patients. This research demonstrates that RepliDec and RepliDec+ can accurately predict phage replication cycles and highlights the critical role of temperate phages in health. Additionally, I investigate the presence of temperate phages within phage cocktails using RepliDec and successfully identify a group of temperate phage sequences. This poses a potentially challenging issue in selecting phage in phage cocktails. Next, I identify multiple phage-bacteria association pairs in bulk using a modified viral tagging (VT) technique and a computational workflow. This method may simplify the future characteristics of phage-bacteria pairs by bypassing culturing, which is especially useful for choosing phages that infect specific bacterial hosts for phage cocktails. Additionally, RepliDec and RepliDec+ could assist in the preliminary characterization of phages to avoid the presence of temperate phages in phage cocktails.

However, these are just the tip of the iceberg, indicating that there is much more beneath the surface that we have yet to explore or understand. Although many databases and tools can be used to study the phage-bacteria interaction and virome data, there is a lot of “dark matter” that can not be solved using these available sources (Santiago-Rodriguez & Hollister, 2022). New sequencing data is generated daily, making it essential to regularly update these databases. The development of novel computational tools relies on new data or advanced algorithms for analyzing viromic data and studying interactions, which will be crucial for enhancing our understanding of the role of phages in health and facilitating their extensive applications.

5. Conclusions and outlook

This work presents a tool, RepliDec, along with an integrated pipeline, RepliDec+, aimed at improving the lifecycle prediction of phages. RepliDec outperforms other tools in the analysis of complete phage genomes as well as simulated metagenomic contigs. This could significantly facilitate a better prediction of phage lifecycle, enabling researchers to understand their behavior, interactions, and ecological roles more effectively. Using RepliDec+, I investigate the impact of temperate phages on patients with IBD, revealing that individuals with severe symptoms exhibit a higher abundance of temperate phages compared to healthy controls. This finding could help the treatment of patients suffering from IBD. Healthcare professionals may develop more effective strategies to improve patient outcomes and enhance the overall management of the condition. Although RepliDec is capable of predicting three types of lifecycle, pseudolysogeny cannot be predicted by any existing tools. This type of lifecycle should also be incorporated into future updates of RepliDec. Additionally, analyzing the genes involved in the phage replication cycle is essential for fully understanding these processes and may yield insights for future applications of phages, particularly in phage therapy

This work also detected the presence of temperate phages in the PYO cocktail. This finding highlights the need for a more careful selection process of phages when formulating a phage cocktail. RepliDec and RepliDec+ can be used as preliminary tools to assess the phage lifecycle, ensuring that more appropriate phages are selected. Additionally, the introduction of temperate phages could have unforeseen effects on the efficacy and safety of treatments for patients. There is very little research about the specific effects temperate phages may have on patients receiving these cocktails. Therefore, it is crucial to conduct further investigations to understand these implications and reach a clear conclusion about their role in therapeutic applications.

This study utilizes a modified viral tagging (VT) method to identify phage-bacteria associations in both patients and healthy controls, setting the stage for future research aimed at a more comprehensive exploration of phage-bacteria interaction pairs. However, VT technology can only confirm the attachment of phages to bacteria. Whether they will interact and how they will interact is unknown; therefore, there is a necessity for further experimental and computational studies to clarify this aspect. Additionally, the phage-bacteria interactions can facilitate the selection of phages for cocktail formulations. Future efforts should emphasize isolating and characterizing these phages, and RepliDec and RepliDec+ could assist in the preliminary characterization of phages to avoid the presence of temperate phages in phage cocktails.

References

Abedon, S. T. (2022). Prophages Preventing Phage Superinfection. In S. T. Abedon, *Bacteriophages as Drivers of Evolution* (pp. 179–191). Springer International Publishing. https://doi.org/10.1007/978-3-030-94309-7_16

Ackermann, H.-W. (2007). 5500 Phages examined in the electron microscope. *Archives of Virology*, *152*(2), 227–243. https://doi.org/10.1007/s00705-006-0849-1

Ahishali, E., Pinarbasi, B., Akyuz, F., Ibrisim, D., Kaymakoglu, S., & Mungan, Z. (2007). A case of Aeromonas hydrophila enteritis in the course of ulcerative colitis. *European Journal of Internal Medicine*, *18*(5), 430–431. https://doi.org/10.1016/j.ejim.2006.12.008

Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). PhiSpy: A novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. *Nucleic Acids Research*, *40*(16), e126–e126. https://doi.org/10.1093/nar/gks406

Al-Bayati, L., Nayeri Fasaei, B., Merat, S., Bahonar, A., & Ghoddusi, A. (2023). Quantitative analysis of the three gut microbiota in UC and non-UC patients using real-time PCR. *Microbial Pathogenesis*, *181*, 106198. https://doi.org/10.1016/j.micpath.2023.106198

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., Baker, S. J. C., Dave, M., McCarthy, M. C., Mukiri, K. M., Nasir, J. A., Golbon, B., Imtiaz, H., Jiang, X., Kaur, K., … McArthur, A. G. (2023). CARD 2023: Expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Research*, *51*(D1), D690–D699. https://doi.org/10.1093/nar/gkac920

Alikhan, N.-F., Petty, N. K., Ben Zakour, N. L., & Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): Simple prokaryote genome comparisons. *BMC Genomics*, *12*(1), 402. https://doi.org/10.1186/1471-2164-12-402

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z. J., Pollard, K. S., Sakharova, E., Parks, D. H., Hugenholtz, P., Segata, N., Kyrpides, N. C., & Finn, R. D. (2021). A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nature Biotechnology*, *39*(1), 105–114. https://doi.org/10.1038/s41587-020-0603-3

Al-Shayeb, B., Sachdeva, R., Chen, L.-X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Méheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., … Banfield, J. F. (2020). Clades of huge phages from across Earth’s ecosystems. *Nature*, *578*(7795), 425–431. https://doi.org/10.1038/s41586-020-2007-4

Amber, T. (2024, July 6). Ulcerative Colitis vs. Crohn’s Disease: What’s the Difference? *Verywellhealth*. https://www.verywellhealth.com/understand-the-differences-between-cd-and-uc-1943108

Amos, G. C. A., Sergaki, C., Logan, A., Iriarte, R., Bannaga, A., Chandrapalan, S., Wellington, E. M. H., Rijpkema, S., & Arasaradnam, R. P. (2021). Exploring how microbiome signatures change across inflammatory bowel disease conditions and disease locations. *Scientific Reports*, *11*(1), 18699. https://doi.org/10.1038/s41598-021-96942-z

Aslam, S., Courtwright, A. M., Koval, C., Lehman, S. M., Morales, S., Furr, C.-L. L., Rosas, F., Brownstein, M. J., Fackler, J. R., Sisson, B. M., Biswas, B., Henry, M., Luu, T., Bivens, B. N., Hamilton, T., Duplessis, C., Logan, C., Law, N., Yung, G., … Schooley, R. T.

(2019). Early clinical experience of bacteriophage therapy in 3 lung transplant recipients. *American Journal of Transplantation*, *19*(9), 2631–2639. https://doi.org/10.1111/ajt.15503

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48. https://doi.org/10.1093/nar/28.1.45

Basha, O. M., Hafez, R. A., Salem, S. M., Anis, R. H., & Hanafy, A. S. (2022). Impact of gut Microbiome alteration in Ulcerative Colitis patients on disease severity and outcome. *Clinical and Experimental Medicine*, *23*(5), 1763–1772. https://doi.org/10.1007/s10238-022-00917-x

Bochow, S., Elliman, J., & Owens, L. (2012). Bacteriophage adenine methyltransferase: A life cycle regulator? Modelled using *Vibrio harveyi* myovirus like. *Journal of Applied Microbiology*, *113*(5), 1001–1013. https://doi.org/10.1111/j.1365-2672.2012.05358.x

Breitbart, M., Thompson, L., Suttle, C., & Sullivan, M. (2007). Exploring the Vast Diversity of Marine Viruses. *Oceanography*, *20*(2), 135–139. https://doi.org/10.5670/oceanog.2007.58

Brister, J. R., Ako-adjei, D., Bao, Y., & Blinkova, O. (2015). NCBI Viral Genomes Resource. *Nucleic Acids Research*, *43*(D1), D571–D577. https://doi.org/10.1093/nar/gku1207

Bull, J. J., & Gill, J. J. (2014). The habits of highly effective phages: Population dynamics as a framework for identifying therapeutic phages. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00618

Callanan, J., Stockdale, S. R., Adriaenssens, E. M., Kuhn, J. H., Rumnieks, J., Pallen, M. J., Shkoporov, A. N., Draper, L. A., Ross, R. P., & Hill, C. (2021). Leviviricetes: Expanding and restructuring the taxonomy of bacteria-infecting single-stranded RNA viruses. *Microbial Genomics*, *7*(11). https://doi.org/10.1099/mgen.0.000686

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S. G., Nayfach, S., & Kyrpides, N. C. (2024). Identification of mobile genetic elements with geNomad. *Nature Biotechnology*, *42*(8), 1303–1312. https://doi.org/10.1038/s41587-023-01953-y

Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D., & Lawley, T. D. (2021). Massive expansion of human gut bacteriophage diversity. *Cell*, *184*(4), 1098-1109.e9. https://doi.org/10.1016/j.cell.2021.01.029

Cao, Y., Shen, J., & Ran, Z. H. (2014). Association between *Faecalibacterium prausnitzii* Reduction and Inflammatory Bowel Disease: A Meta-Analysis and Systematic Review of the Literature. *Gastroenterology Research and Practice*, *2014*, 1–7. https://doi.org/10.1155/2014/872725

Cao, Z., Sugimura, N., Burgermeister, E., Ebert, M. P., Zuo, T., & Lan, P. (2022). The gut virome: A new microbiome component in health and disease. *eBioMedicine*, *81*, 104113. https://doi.org/10.1016/j.ebiom.2022.104113

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics (Oxford, England)*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Champagne-Jorgensen, K., Luong, T., Darby, T., & Roach, D. R. (2023). Immunogenicity of bacteriophages. *Trends in Microbiology*, *31*(10), 1058–1071. https://doi.org/10.1016/j.tim.2023.04.008

Chanishvili, N. (2012). Phage Therapy—History from Twort and d'Herelle Through Soviet Experience to Current Approaches. In *Advances in Virus Research* (Vol. 83, pp. 3–40). Elsevier. https://doi.org/10.1016/B978-0-12-394438-2.00001-3

Chattopadhyay, I., Dhar, R., Pethusamy, K., Seethy, A., Srivastava, T., Sah, R., Sharma, J., & Karmakar, S. (2021). Exploring the Role of Gut Microbiome in Colon Cancer. *Applied Biochemistry and Biotechnology*, *193*(6), 1780–1799. https://doi.org/10.1007/s12010-021-03498-9

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, *34*(17), i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Chen, X., Weinbauer, M. G., Jiao, N., & Zhang, R. (2021). Revisiting marine lytic and lysogenic virus-host interactions: Kill-the-Winner and Piggyback-the-Winner. *Science Bulletin*, *66*(9), 871–874. https://doi.org/10.1016/j.scib.2020.12.014

Clokie, M. R., Millard, A. D., Letarov, A. V., & Heaphy, S. (2011). Phages in nature. *Bacteriophage*, *1*(1), 31–45. PubMed. https://doi.org/10.4161/bact.1.1.14942

Clooney, A. G., Sutton, T. D. S., Shkoporov, A. N., Holohan, R. K., Daly, K. M., O'Regan, O., Ryan, F. J., Draper, L. A., Plevy, S. E., Ross, R. P., & Hill, C. (2019). Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host & Microbe*, *26*(6), 764-778.e5. https://doi.org/10.1016/j.chom.2019.10.009

Colavecchio, A., D'Souza, Y., Tompkins, E., Jeukens, J., Freschi, L., Emond-Rheault, J.-G., Kukavica-Ibrulj, I., Boyle, B., Bekal, S., Tamber, S., Levesque, R. C., & Goodridge, L. D. (2017). Prophage Integrase Typing Is a Useful Indicator of Genomic Diversity in Salmonella enterica. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01283

Colorectal Cancer Statistics | CDC. (2022, November 29). https://www.cdc.gov/cancer/colorectal/statistics/index.htm

Colorectal_cancer_factsheet-Mar_2021.pdf. (n.d.). Retrieved April 20, 2023, from https://ecis.jrc.ec.europa.eu/pdf/Colorectal_cancer_factsheet-Mar_2021.pdf

Comeau, A. M., Hatfull, G. F., Krisch, H. M., Lindell, D., Mann, N. H., & Prangishvili, D. (2008). Exploring the prokaryotic virosphere. *Research in Microbiology*, *159*(5), 306–313.

Costa, S. P., Cunha, A. P., Freitas, P. P., & Carvalho, C. M. (2022). A Phage Receptor-Binding Protein as a Promising Tool for the Detection of Escherichia coli in Human Specimens. *Frontiers in Microbiology*, *13*, 871855. https://doi.org/10.3389/fmicb.2022.871855

d'Herelle, M. F. (1961). Sur un microbe invisible antagoniste des bacilles dysentériques. *Acta Kravsi*.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

Danovaro, R., Corinaldesi, C., Dell'Anno, A., Fuhrman, J. A., Middelburg, J. J., Noble, R. T., & Suttle, C. A. (2011). Marine viruses and global climate change. *FEMS Microbiology Reviews*, *35*(6), 993–1034. https://doi.org/10.1111/j.1574-6976.2010.00258.x

Davies, E. V., James, C. E., Williams, D., O'Brien, S., Fothergill, J. L., Haldenby, S., Paterson, S., Winstanley, C., & Brockhurst, M. A. (2016). Temperate phages both mediate and drive adaptive evolution in pathogen biofilms. *Proceedings of the National Academy of Sciences*, *113*(29), 8266–8271. https://doi.org/10.1073/pnas.1520056113

De Caro, G., Gaiani, F., Duranti, S., Fugazza, A., Madia, C., Milani, C., Mancabelli, L., Turroni, F., de' Angelis, G. L., Carra, M. C., de' Angelis, N., & Ventura, M. (2016). The Role of Bifidobacteria in Ulcerative Colitis: Preliminary Results: 723. *Official Journal of the American College of Gastroenterology | ACG*, *111*.

https://journals.lww.com/ajg/fulltext/2016/10001/the_role_of_bifidobacteria_in_ulcerative_colitis_.723.aspx

De Melo, A. C. C., Da Mata Gomes, A., Melo, F. L., Ardisson-Araújo, D. M. P., De Vargas, A. P. C., Ely, V. L., Kitajima, E. W., Ribeiro, B. M., & Wolff, J. L. C. (2019). Characterization of a bacteriophage with broad host range against strains of Pseudomonas aeruginosa isolated from domestic animals. *BMC Microbiology*, *19*(1), 134. https://doi.org/10.1186/s12866-019-1481-z

Dedrick, R. M., Mavrich, T. N., Ng, W. L., Cervantes Reyes, J. C., Olm, M. R., Rush, R. E., Jacobs-Sera, D., Russell, D. A., & Hatfull, G. F. (2016). Function, expression, specificity, diversity and incompatibility of actinobacteriophage *parABS* systems. *Molecular Microbiology*, *101*(4), 625–644. https://doi.org/10.1111/mmi.13414

Deng, L., Ignacio-Espinoza, J. C., Gregory, A. C., Poulos, B. T., Weitz, J. S., Hugenholtz, P., & Sullivan, M. B. (2014). Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature*, *513*(7517), 242–245. https://doi.org/10.1038/nature13459

Diard, M., Bakkeren, E., Cornuault, J. K., Moor, K., Hausmann, A., Sellin, M. E., Loverdo, C., Aertsen, A., Ackermann, M., De Paepe, M., Slack, E., & Hardt, W.-D. (2017). Inflammation boosts bacteriophage transfer between *Salmonella* spp. *Science*, *355*(6330), 1211–1215. https://doi.org/10.1126/science.aaf8451

Dion, M. B., Plante, P.-L., Zufferey, E., Shah, S. A., Corbeil, J., & Moineau, S. (2021). Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Research*, *49*(6), 3127–3138. https://doi.org/10.1093/nar/gkab133

Duerkop, B. A., Kleiner, M., Paez-Espino, D., Zhu, W., Bushnell, B., Hassell, B., Winter, S. E., Kyrpides, N. C., & Hooper, L. V. (2018). Murine colitis reveals a disease-associated bacteriophage community. *Nature Microbiology*, *3*(9), 1023–1031. https://doi.org/10.1038/s41564-018-0210-y

Duranti, S., Gaiani, F., Mancabelli, L., Milani, C., Grandi, A., Bolchi, A., Santoni, A., Lugli, G. A., Ferrario, C., Mangifesta, M., Viappiani, A., Bertoni, S., Vivo, V., Serafini, F., Barbaro, M. R., Fugazza, A., Barbara, G., Gioiosa, L., Palanza, P., … Turroni, F. (2016). Elucidating the gut microbiome of ulcerative colitis: Bifidobacteria as novel microbial biomarkers. *FEMS Microbiology Ecology*, *92*(12), fiw191. https://doi.org/10.1093/femsec/fiw191

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., Barr, J. J., Speth, D. R., Seguritan, V., Aziz, R. K., Felts, B., Dinsdale, E. A., Mokili, J. L., & Edwards, R. A. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature Communications*, *5*(1), 4498. https://doi.org/10.1038/ncomms5498

Džunková, M., Low, S. J., Daly, J. N., Deng, L., Rinke, C., & Hugenholtz, P. (2019). Defining the human gut host–phage network through single-cell viral tagging. *Nature Microbiology*, *4*(12), 2192–2203. https://doi.org/10.1038/s41564-019-0526-2

E. White, H., & V. Orlova, E. (2020). Bacteriophages: Their Structural Organisation and Function. In R. Savva (Ed.), *Bacteriophages—Perspectives and Future*. IntechOpen. https://doi.org/10.5772/intechopen.85484

Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiology Reviews*, *40*(2), 258–272. https://doi.org/10.1093/femsre/fuv048

Elfadadny, A., Ragab, R. F., Abou Shehata, M. A., Elfadadny, M. R., Farag, A., Abd El-Aziz, A. H., & Khalifa, H. O. (2024). Exploring Bacteriophage Applications in Medicine and

Beyond. *Acta Microbiologica Hellenica*, *69*(3), 167–179. https://doi.org/10.3390/amh69030016

Erez, Z., Steinberger-Levy, I., Shamir, M., Doron, S., Stokar-Avihail, A., Peleg, Y., Melamed, S., Leavitt, A., Savidor, A., Albeck, S., Amitai, G., & Sorek, R. (2017). Communication between viruses guides lysis–lysogeny decisions. *Nature*, *541*(7638), 488–493. https://doi.org/10.1038/nature21049

European Federation of Crohn's and Ulcerative Colitis Associations (EFCCA). (n.d). *What is IBD?* https://efcca.org/content/what-ibd

Fang, C.-Y., Chen, J.-S., Hsu, B.-M., Hussain, B., Rathod, J., & Lee, K.-H. (2021). Colorectal Cancer Stage-Specific Fecal Bacterial Community Fingerprinting of the Taiwanese Population and Underpinning of Potential Taxonomic Biomarkers. *Microorganisms*, *9*(8), 1548. https://doi.org/10.3390/microorganisms9081548

Fang, Z., Tan, J., Wu, S., Li, M., Xu, C., Xie, Z., & Zhu, H. (2019). PPR-Meta: A tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, *8*(6), giz066. https://doi.org/10.1093/gigascience/giz066

Farb, J. N., & Morrical, S. W. (2009). Functional complementation of UvsX and UvsY mutations in the mediation of T4 homologous recombination. *Nucleic Acids Research*, *37*(7), 2336–2345. https://doi.org/10.1093/nar/gkp096

Federici, S., Kredo-Russo, S., Valdés-Mas, R., Kviatcovsky, D., Weinstock, E., Matiuhin, Y., Silberberg, Y., Atarashi, K., Furuichi, M., Oka, A., Liu, B., Fibelman, M., Weiner, I. N., Khabra, E., Cullin, N., Ben-Yishai, N., Inbar, D., Ben-David, H., Nicenboim, J., … Elinav, E. (2022). Targeted suppression of human IBD-associated gut microbiota commensals by phage consortia for treatment of intestinal inflammation. *Cell*, *185*(16), 2879-2898.e24. https://doi.org/10.1016/j.cell.2022.07.003

Focardi, A., Ostrowski, M., Goossen, K., Brown, M. V., & Paulsen, I. (2020). Investigating the Diversity of Marine Bacteriophage in Contrasting Water Masses Associated with the East Australian Current (EAC) System. *Viruses*, *12*(3), 317. https://doi.org/10.3390/v12030317

Fong, K., Wong, C. W. Y., Wang, S., & Delaquis, P. (2021). How Broad Is Enough: The Host Range of Bacteriophages and Its Impact on the Agri-Food Sector. *PHAGE*, *2*(2), 83–91. https://doi.org/10.1089/phage.2020.0036

Fowoyo, P. T. (2024). Phage Therapy: Clinical Applications, Efficacy, and Implementation Hurdles. *The Open Microbiology Journal*, *18*(1), e18742858281566. https://doi.org/10.2174/0118742858281566231221045303

Gajewski, S., Webb, M. R., Galkin, V., Egelman, E. H., Kreuzer, K. N., & White, S. W. (2011). Crystal Structure of the Phage T4 Recombinase UvsX and Its Functional Interaction with the T4 SF2 Helicase UvsW. *Journal of Molecular Biology*, *405*(1), 65–76. https://doi.org/10.1016/j.jmb.2010.10.004

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, *49*(D1), D274–D281. https://doi.org/10.1093/nar/gkaa1018

Gao, R., Zhu, Y., Kong, C., Xia, K., Li, H., Zhu, Y., Zhang, X., Liu, Y., Zhong, H., Yang, R., Chen, C., Qin, N., & Qin, H. (2021). Alterations, Interactions, and Diagnostic Potential of Gut Bacteria and Viruses in Colorectal Cancer. *Frontiers in Cellular and Infection Microbiology*, *11*, 657867. https://doi.org/10.3389/fcimb.2021.657867

García-Cruz, J. C., Huelgas-Méndez, D., Jiménez-Zúñiga, J. S., Rebollar-Juárez, X., Hernández-Garnica, M., Fernández-Presas, A. M., Husain, F. M., Alenazy, R., Alqasmi, M., Albalawi, T., Alam, P., & García-Contreras, R. (2023). Myriad applications of bacteriophages beyond phage therapy. *PeerJ*, *11*, e15272. https://doi.org/10.7717/peerj.15272

Georjon, H., & Bernheim, A. (2023). The highly diverse antiphage defence systems of bacteria. *Nature Reviews Microbiology*, *21*(10), 686–700. https://doi.org/10.1038/s41579-023-00934-x

Gohil, K., & Carramusa, B. (2014). Ulcerative colitis and Crohn's disease. *P & T: A Peer-Reviewed Journal for Formulary Management*, *39*(8), 576–577.

Göller, P. C., Elsener, T., Lorgé, D., Radulovic, N., Bernardi, V., Naumann, A., Amri, N., Khatchatourova, E., Coutinho, F. H., Loessner, M. J., & Gómez-Sanz, E. (2021). Multi-species host range of staphylococcal phages isolated from wastewater. *Nature Communications*, *12*(1), 6965. https://doi.org/10.1038/s41467-021-27037-6

Grazziotin, A. L., Koonin, E. V., & Kristensen, D. M. (2017). Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation. *Nucleic Acids Research*, *45*(D1), D491–D498. https://doi.org/10.1093/nar/gkw975

Grigoriev, A. (1998). Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research*, *26*(10), 2286–2290. https://doi.org/10.1093/nar/26.10.2286

Guo, J., Bolduc, B., Zayed, A. A., Varsani, A., Dominguez-Huerta, G., Delmont, T. O., Pratama, A. A., Gazitúa, M. C., Vik, D., Sullivan, M. B., & Roux, S. (2021). VirSorter2: A multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*, *9*(1), 37. https://doi.org/10.1186/s40168-020-00990-y

Gupta, P., Singh, H. S., Shukla, V. K., Nath, G., & Bhartiya, S. K. (2019). Bacteriophage Therapy of Chronic Nonhealing Wound: Clinical Study. *The International Journal of Lower Extremity Wounds*, *18*(2), 171–175. https://doi.org/10.1177/1534734619835115

Hamilton, A. L., Kamm, M. A., Ng, S. C., & Morrison, M. (2018). Proteus spp. As Putative Gastrointestinal Pathogens. *Clinical Microbiology Reviews*, *31*(3), e00085-17. https://doi.org/10.1128/CMR.00085-17

Han, M., Yang, P., Zhong, C., & Ning, K. (2018). The Human Gut Virome in Hypertension. *Frontiers in Microbiology*, *9*, 3150. https://doi.org/10.3389/fmicb.2018.03150

Hatfull, G. F. (2008). Bacteriophage genomics. *Current Opinion in Microbiology*, *11*(5), 447–453. https://doi.org/10.1016/j.mib.2008.09.004

Hatfull, G. F., & Hendrix, R. W. (2011). Bacteriophages and their genomes. *Current Opinion in Virology*, *1*(4), 298–303. https://doi.org/10.1016/j.coviro.2011.06.009

Hay, I. D., & Lithgow, T. (2019). Filamentous phages: Masters of a microbial sharing economy. *EMBO Reports*, *20*(6), e47427. https://doi.org/10.15252/embr.201847427

Henrot, C., & Petit, M. (2022). Signals triggering prophage induction in the gut microbiota. *Molecular Microbiology*, *118*(5), 494–502. https://doi.org/10.1111/mmi.14983

Hesketh-Best, P. J., Bosco-Santos, A., Garcia, S. L., O'Beirne, M. D., Werne, J. P., Gilhooly, W. P., & Silveira, C. B. (2023). Viruses of sulfur oxidizing phototrophs encode genes for pigment, carbon, and sulfur metabolisms. *Communications Earth & Environment*, *4*(1), 126. https://doi.org/10.1038/s43247-023-00796-4

Hitchcock, N. M., Devequi Gomes Nunes, D., Shiach, J., Valeria Saraiva Hodel, K., Dantas Viana Barbosa, J., Alencar Pereira Rodrigues, L., Coler, B. S., Botelho Pereira Soares, M., & Badaró, R. (2023). Current Clinical Landscape and Global Potential of Bacteriophage Therapy. *Viruses*, *15*(4), 1020. https://doi.org/10.3390/v15041020

Hobbs, Z., & Abedon, S. T. (2016). Diversity of phage infection types and associated terminology: The problem with ‘Lytic or lysogenic.’ *FEMS Microbiology Letters*, *363*(7), fnw047. https://doi.org/10.1093/femsle/fnw047

Hockenberry, A. J., & Wilke, C. O. (2021). *BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains*. 6.

Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Frontiers in Microbiology*, *11*, 582779. https://doi.org/10.3389/fmicb.2020.582779

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594. https://doi.org/10.1093/bioinformatics/btr708

Hurwitz, B. L., & U’Ren, J. M. (2016). Viral metabolic reprogramming in marine ecosystems. *Current Opinion in Microbiology*, *31*, 161–168. https://doi.org/10.1016/j.mib.2016.04.002

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*(1), 119. https://doi.org/10.1186/1471-2105-11-119

Irshath, A. A., Rajan, A. P., Vimal, S., Prabhakaran, V.-S., & Ganesan, R. (2023). Bacterial Pathogenesis in Various Fish Diseases: Recent Advances and Specific Challenges in Vaccine Development. *Vaccines*, *11*(2), 470. https://doi.org/10.3390/vaccines11020470

Islam, Md. S., Nime, I., Pan, F., & Wang, X. (2023). Isolation and characterization of phage ISTP3 for bio-control application against drug-resistant Salmonella. *Frontiers in Microbiology*, *14*, 1260181. https://doi.org/10.3389/fmicb.2023.1260181

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1), 5114. https://doi.org/10.1038/s41467-018-07641-9

Jault, P., Leclerc, T., Jennes, S., Pirnay, J. P., Que, Y.-A., Resch, G., Rousseau, A. F., Ravat, F., Carsin, H., Le Floch, R., Schaal, J. V., Soler, C., Fevre, C., Arnaud, I., Bretaudeau, L., & Gabard, J. (2019). Efficacy and tolerability of a cocktail of bacteriophages to treat burn wounds infected by Pseudomonas aeruginosa (PhagoBurn): A randomised, controlled, double-blind phase 1/2 trial. *The Lancet Infectious Diseases*, *19*(1), 35–45. https://doi.org/10.1016/S1473-3099(18)30482-1

Jędrusiak, A., Fortuna, W., Majewska, J., Górski, A., & Jończyk-Matysiak, E. (2023). Phage Interactions with the Nervous System in Health and Disease. *Cells*, *12*(13), 1720. https://doi.org/10.3390/cells12131720

Joachim, J. (2022, February 21). *Microbiome analysis of viruses is more accessible than ever*. https://communities.springernature.com/posts/microbiome-analysis-of-viruses-is-more-accessible-than-ever

Johansen, J., Plichta, D. R., Nissen, J. N., Jespersen, M. L., Shah, S. A., Deng, L., Stokholm, J., Bisgaard, H., Nielsen, D. S., Sørensen, S. J., & Rasmussen, S. (2022). Genome binning of viral entities from bulk metagenomics data. *Nature Communications*, *13*(1), 965. https://doi.org/10.1038/s41467-022-28581-5

Johnson, J. S., Spakowicz, D. J., Hong, B.-Y., Petersen, L. M., Demkowicz, P., Chen, L., Leopold, S. R., Hanson, B. M., Agresta, H. O., Gerstein, M., Sodergren, E., & Weinstock, G. M. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nature Communications*, *10*(1), 5029. https://doi.org/10.1038/s41467-019-13036-1

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(suppl_2), W5–W9. https://doi.org/10.1093/nar/gkn201

Jurczak-Kurek, A., Gąsior, T., Nejman-Faleńczyk, B., Bloch, S., Dydecka, A., Topka, G., Necel, A., Jakubowska-Deredas, M., Narajczyk, M., Richert, M., Mieszkowska, A., Wróbel, B., Węgrzyn, G., & Węgrzyn, A. (2016). Biodiversity of bacteriophages: Morphological and biological properties of a large group of phages isolated from urban sewage. *Scientific Reports*, *6*(1), 34338. https://doi.org/10.1038/srep34338

Jurtz, V. I., Villarroel, J., Lund, O., Voldby Larsen, M., & Nielsen, M. (2016). MetaPhinder—Identifying bacteriophage sequences in metagenomic data sets. *PLoS One*, *11*(9), e0163111.

Kanareykina, S. K., Misautova, A. A., Zlatkina, A. R., & Levina, E. N. (1987). Proteus dysbioses in patients with ulcerative colitis. *Food / Nahrung*, *31*(5–6), 557–561. https://doi.org/10.1002/food.19870310570

Kanehisa, M., & Goto, S. (2000). *KEGG: Kyoto Encyclopedia of Genes and Genomes*. 4.

Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *7*, e7359. https://doi.org/10.7717/peerj.7359

Kang, D.-Y., Park, J.-L., Yeo, M.-K., Kang, S.-B., Kim, J.-M., Kim, J. S., & Kim, S.-Y. (2023). Diagnosis of Crohn’s disease and ulcerative colitis using the microbiome. *BMC Microbiology*, *23*(1), 336. https://doi.org/10.1186/s12866-023-03084-5

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Keen, E. C. (2015). A century of phage research: Bacteriophages and the shaping of modern biology. *BioEssays*, *37*(1), 6–9. https://doi.org/10.1002/bies.201400152

Kennedy, J. M., De Silva, A., Walton, G. E., & Gibson, G. R. (2024). A review on the use of prebiotics in ulcerative colitis. *Trends in Microbiology*, *32*(5), 507–515. https://doi.org/10.1016/j.tim.2023.11.007

Khan Mirzaei, M., Khan, Md. A. A., Ghosh, P., Taranu, Z. E., Taguer, M., Ru, J., Chowdhury, R., Kabir, Md. M., Deng, L., Mondal, D., & Maurice, C. F. (2020). Bacteriophages Isolated from Stunted Children Can Regulate Gut Bacterial Communities in an Age-Specific Manner. *Cell Host & Microbe*, *27*(2), 199-212.e5. https://doi.org/10.1016/j.chom.2020.01.004

Khan Mirzaei, M., Xue, J., Costa, R., Ru, J., Schulz, S., Taranu, Z. E., & Deng, L. (2021). Challenges of Studying the Human Virome – Relevant Emerging Technologies. *Trends in Microbiology*, *29*(2), 171–181. https://doi.org/10.1016/j.tim.2020.05.021

Kieft, K., Adams, A., Salamzade, R., Kalan, L., & Anantharaman, K. (2022). vRhyme enables binning of viral genomes from metagenomes. *Nucleic Acids Research*, *50*(14), e83–e83. https://doi.org/10.1093/nar/gkac341

Kieft, K., Breister, A. M., Huss, P., Linz, A. M., Zanetakos, E., Zhou, Z., Rahlff, J., Esser, S. P., Probst, A. J., Raman, S., Roux, S., & Anantharaman, K. (2021). Virus-associated organosulfur metabolism in human and environmental systems. *Cell Reports*, *36*(5), 109471. https://doi.org/10.1016/j.celrep.2021.109471

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*(1), 90. https://doi.org/10.1186/s40168-020-00867-0

Kieft, K., Zhou, Z., Anderson, R. E., Buchan, A., Campbell, B. J., Hallam, S. J., Hess, M., Sullivan, M. B., Walsh, D. A., Roux, S., & Anantharaman, K. (2021). Ecology of inorganic sulfur auxiliary metabolism in widespread bacteriophages. *Nature Communications*, *12*(1), 3503. https://doi.org/10.1038/s41467-021-23698-5

Knezevic, P., Adriaenssens, E. M., & ICTV Report Consortium. (2021). ICTV Virus Taxonomy Profile: Inoviridae. *Journal of General Virology*, *102*(7). https://doi.org/10.1099/jgv.0.001614

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L.-I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R., Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Koonin, E. V., Dolja, V. V., Krupovic, M., Varsani, A., Wolf, Y. I., Yutin, N., Zerbini, F. M., & Kuhn, J. H. (2020). Global Organization and Proposed Megataxonomy of the Virus World. *Microbiology and Molecular Biology Reviews*, *84*(2), e00061-19. https://doi.org/10.1128/MMBR.00061-19

Kuipers, S., Ruth, M. M., Mientjes, M., De Sévaux, R. G. L., & Van Ingen, J. (2019). A Dutch Case Report of Successful Treatment of Chronic Relapsing Urinary Tract Infection with Bacteriophages in a Renal Transplant Patient. *Antimicrobial Agents and Chemotherapy*, *64*(1), e01281-19. https://doi.org/10.1128/AAC.01281-19

Kutter, E., De Vos, D., Gvasalia, G., Alavidze, Z., Gogokhia, L., Kuhl, S., & Abedon, S. (2010). Phage Therapy in Clinical Practice: Treatment of Human Infections. *Current Pharmaceutical Biotechnology*, *11*(1), 69–86. https://doi.org/10.2174/138920110790725401

Kwan, T., Liu, J., DuBow, M., Gros, P., & Pelletier, J. (2005). The complete genomes and proteomes of 27 *Staphylococcus aureus* bacteriophages. *Proceedings of the National Academy of Sciences*, *102*(14), 5174–5179. https://doi.org/10.1073/pnas.0501140102

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

LaVergne, S., Hamilton, T., Biswas, B., Kumaraswamy, M., Schooley, R. T., & Wooten, D. (2018). Phage Therapy for a Multidrug-Resistant Acinetobacter baumannii Craniectomy Site Infection. *Open Forum Infectious Diseases*, *5*(4), ofy064. https://doi.org/10.1093/ofid/ofy064

Le Berre, C., Honap, S., & Peyrin-Biroulet, L. (2023). Ulcerative colitis. *The Lancet*, *402*(10401), 571–584. https://doi.org/10.1016/S0140-6736(23)00966-2

Lebeaux, D., Merabishvili, M., Caudron, E., Lannoy, D., Van Simaey, L., Duyvejonck, H., Guillemain, R., Thumerelle, C., Podglajen, I., Compain, F., Kassis, N., Mainardi, J.-L., Wittmann, J., Rohde, C., Pirnay, J.-P., Dufour, N., Vermeulen, S., Gansemans, Y., Van Nieuwerburgh, F., & Vaneechoutte, M. (2021). A Case of Phage Therapy against Pandrug-Resistant Achromobacter xylosoxidans in a 12-Year-Old Lung-Transplanted Cystic Fibrosis Patient. *Viruses*, *13*(1), 60. https://doi.org/10.3390/v13010060

Lee, S., de Bie, B., Ngo, J., & Lo, J. (2018). *O16 Antigen Confers Resistance to Bacteriophage T4 and T7 But Does Not Reduce T4/T7 Adsorption in Escherichia coli K-12. 22*.

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/nar/gkab301

Lewis, J. D., Parlett, L. E., Jonsson Funk, M. L., Brensinger, C., Pate, V., Wu, Q., Dawwas, G. K., Weiss, A., Constant, B. D., McCauley, M., Haynes, K., Yang, J. Y., Schaubel, D. E., Hurtado-Lorenzo, A., & Kappelman, M. D. (2023). Incidence, Prevalence, and Racial and Ethnic Distribution of Inflammatory Bowel Disease in the United States. *Gastroenterology*, *165*(5), 1197-1205.e2. https://doi.org/10.1053/j.gastro.2023.07.003

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics*, *20*(12), 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Sharp, R., Rutherford, K., Gupta, K., & Van Duyne, G. D. (2018). Serine Integrase attP Binding and Specificity. *Journal of Molecular Biology*, *430*(21), 4401–4418. https://doi.org/10.1016/j.jmb.2018.09.007

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Li, X., Chen, Y., Wang, S., Duan, X., Zhang, F., Guo, A., Tao, P., Chen, H., Li, X., & Qian, P. (2022). Exploring the Benefits of Metal Ions in Phage Cocktail for the Treatment of Methicillin-Resistant Staphylococcus aureus (MRSA) Infection. *Infection and Drug Resistance*, *Volume 15*, 2689–2702. https://doi.org/10.2147/IDR.S362743

Liang, X., & Radosevich, M. (2020). Phage Communication and the Ecological Implications on Microbial Interactions, Diversity, and Function. In G. Witzany (Ed.), *Biocommunication of Phages* (pp. 71–86). Springer International Publishing. https://doi.org/10.1007/978-3-030-45885-0_3

LIMITATIONS OF BACTERIOPHAGE THERAPY. (1931). *JAMA: The Journal of the American Medical Association*, *96*(9), 693. https://doi.org/10.1001/jama.1931.02720350045014

Lin, D. M., Koskella, B., & Lin, H. C. (2017). Phage therapy: An alternative to antibiotics in the age of multi-drug resistance. *World Journal of Gastrointestinal Pharmacology and Therapeutics*, *8*(3), 162. https://doi.org/10.4292/wjgpt.v8.i3.162

Liu, B., & Pop, M. (2009). ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Research*, *37*(Database), D443–D447. https://doi.org/10.1093/nar/gkn656

Llor, C., & Bjerrum, L. (2014). Antimicrobial resistance: Risk associated with antibiotic overuse and initiatives to reduce the problem. *Therapeutic Advances in Drug Safety*, *5*(6), 229–241. https://doi.org/10.1177/2042098614554919

Loh, G., & Blaut, M. (2012). Role of commensal gut bacteria in inflammatory bowel diseases. *Gut Microbes*, *3*(6), 544–555. https://doi.org/10.4161/gmic.22156

Loney, R. E., Delesalle, V. A., Chaudry, B. E., Czerpak, M., Guffey, A. A., Goubet-McCall, L., McCarty, M., Strine, M. S., Tanke, N. T., Vill, A. C., & Krukonis, G. P. (2023). A Novel Subcluster of Closely Related Bacillus Phages with Distinct Tail Fiber/Lysin Gene Combinations. *Viruses*, *15*(11), 2267. https://doi.org/10.3390/v15112267

Łoś, M., & Węgrzyn, G. (2012). Pseudolysogeny. In *Advances in Virus Research* (Vol. 82, pp. 339–349). Elsevier. https://doi.org/10.1016/B978-0-12-394621-8.00019-4

Ma, Y., You, X., Mai, G., Tokuyasu, T., & Liu, C. (2018). A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome*, *6*(1), 24. https://doi.org/10.1186/s40168-018-0410-y

Machiels, K., Joossens, M., Sabino, J., De Preter, V., Arijs, I., Eeckhaut, V., Ballet, V., Claes, K., Van Immerseel, F., Verbeke, K., Ferrante, M., Verhaegen, J., Rutgeerts, P., & Vermeire, S. (2014). A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut*, *63*(8), 1275–1283. https://doi.org/10.1136/gutjnl-2013-304833

Maguire, F., Jia, B., Gray, K. L., Lau, W. Y. V., Beiko, R. G., & Brinkman, F. S. L. (2020). Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. *Microbial Genomics*, *6*(10). https://doi.org/10.1099/mgen.0.000436

Marbouty, M., Thierry, A., Millot, G. A., & Koszul, R. (2021). MetaHiC phage-bacteria infection network reveals active cycling phages of the healthy human gut. *eLife*, *10*, e60608. https://doi.org/10.7554/eLife.60608

Mavrich, T. N., & Hatfull, G. F. (2017). Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology*, *2*(9), 17112. https://doi.org/10.1038/nmicrobiol.2017.112

McDonald, J. (2016). Ocean viruses may have impact on Earth’s climate. *Science*. https://doi.org/10.1126/science.aag0620

McNair, K., Bailey, B. A., & Edwards, R. A. (2012). PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics*, *28*(5), 614–618. https://doi.org/10.1093/bioinformatics/bts014

McNair, K., Zhou, C., Dinsdale, E. A., Souza, B., & Edwards, R. A. (2019). PHANOTATE: A novel approach to gene identification in phage genomes. *Bioinformatics*, *35*(22), 4537–4542. https://doi.org/10.1093/bioinformatics/btz265

Megremis, S., Constantinides, B., Xepapadaki, P., Yap, C. F., Sotiropoulos, A. G., Bachert, C., Finotto, S., Jartti, T., Tapinos, A., Vuorinen, T., Andreakos, E., Robertson, D. L., & Papadopoulos, N. G. (2023). Respiratory eukaryotic virome expansion and bacteriophage deficiency characterize childhood asthma. *Scientific Reports*, *13*(1), 8319. https://doi.org/10.1038/s41598-023-34730-7

Microbiology by numbers. (2011). *Nature Reviews Microbiology*, *9*(9), 628–628. https://doi.org/10.1038/nrmicro2644

Mills, R. H., Dulai, P. S., Vázquez-Baeza, Y., Sauceda, C., Daniel, N., Gerner, R. R., Batachari, L. E., Malfavon, M., Zhu, Q., Weldon, K., Humphrey, G., Carrillo-Terrazas, M., Goldasich, L. D., Bryant, M., Raffatellu, M., Quinn, R. A., Gewirtz, A. T., Chassaing, B., Chu, H., … Gonzalez, D. J. (2022). Multi-omics analyses of the ulcerative colitis gut microbiome link Bacteroides vulgatus proteases with disease severity. *Nature Microbiology*, *7*(2), 262–276. https://doi.org/10.1038/s41564-021-01050-3

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirzaei, M. K., & Maurice, C. F. (2017). Ménage à trois in the human gut: Interactions between host, bacteria and phages. *Nature Reviews Microbiology*, *15*(7), 397–408. https://doi.org/10.1038/nrmicro.2017.30

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121–e121. https://doi.org/10.1093/nar/gkt263

Mitrofanov, A., Alkhnbashi, O. S., Shmakov, S. A., Makarova, K. S., Koonin, E. V., & Backofen, R. (2021). CRISPRidentify: Identification of CRISPR arrays using machine learning approach. *Nucleic Acids Research*, *49*(4), e20–e20. https://doi.org/10.1093/nar/gkaa1158

Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E., & Ghai, R. (2013). Expanding the Marine Virosphere Using Metagenomics. *PLoS Genetics*, *9*(12), e1003987. https://doi.org/10.1371/journal.pgen.1003987

Murchland, I., Ahlgren-Berg, A., Priest, D. G., Dodd, I. B., & Shearwin, K. E. (2014). Promoter Activation by CII, a Potent Transcriptional Activator from Bacteriophage 186. *Journal of Biological Chemistry*, *289*(46), 32094–32108. https://doi.org/10.1074/jbc.M114.608026

Murphy, K. C., Ritchie, J. M., Waldor, M. K., Løbner-Olesen, A., & Marinus, M. G. (2008). Dam Methyltransferase Is Required for Stable Lysogeny of the Shiga Toxin (Stx2)-Encoding Bacteriophage 933W of Enterohemorrhagic *Escherichia coli* O157:H7. *Journal of Bacteriology*, *190*(1), 438–441. https://doi.org/10.1128/JB.01373-07

Murray, C. J. L., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., … Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *The Lancet*, *399*(10325), 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0

Muscatt, G., Hilton, S., Raguideau, S., Teakle, G., Lidbury, I. D. E. A., Wellington, E. M. H., Quince, C., Millard, A., Bending, G. D., & Jameson, E. (2022). Crop management shapes the diversity and activity of DNA and RNA viruses in the rhizosphere. *Microbiome*, *10*(1), 181. https://doi.org/10.1186/s40168-022-01371-3

Nadeem, A., & Wahl, L. M. (2017). Prophage as a genetic reservoir: Promoting diversity and driving innovation in the host community: BRIEF COMMUNICATION. *Evolution*, *71*(8), 2080–2089. https://doi.org/10.1111/evo.13287

Nakatsu, G., Zhou, H., Wu, W. K. K., Wong, S. H., Coker, O. O., Dai, Z., Li, X., Szeto, C.-H., Sugimura, N., Lam, T. Y.-T., Yu, A. C.-S., Wang, X., Chen, Z., Wong, M. C.-S., Ng, S. C., Chan, M. T. V., Chan, P. K. S., Chan, F. K. L., Sung, J. J.-Y., & Yu, J. (2018). Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology*, *155*(2), 529-541.e5. https://doi.org/10.1053/j.gastro.2018.04.018

National Institutes of Health, & US Department of Health. (2009). Opportunities and challenges in digestive diseases research: Recommendations of the national commission on digestive diseases. *Bethesda, MD: National Institutes of Health*.

Nayfach, S., Camargo, A. P., Schulz, F., Eloe-Fadrosh, E., Roux, S., & Kyrpides, N. C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nature Biotechnology*, *39*(5), 578–585. https://doi.org/10.1038/s41587-020-00774-7

Nie, W., Qiu, T., Wei, Y., Ding, H., Guo, Z., & Qiu, J. (2024). Advances in phage–host interaction prediction: *In silico* method enhances the development of phage therapies. *Briefings in Bioinformatics*, *25*(3), bbae117. https://doi.org/10.1093/bib/bbae117

Nishiyama, H., Endo, H., Blanc-Mathieu, R., & Ogata, H. (2020). Ecological Structuring of Temperate Bacteriophages in the Inflammatory Bowel Disease-Affected Gut. *Microorganisms*, *8*(11), 1663. https://doi.org/10.3390/microorganisms8111663

Norman, J. M., Handley, S. A., Baldridge, M. T., Droit, L., Liu, C. Y., Keller, B. C., Kambal, A., Monaco, C. L., Zhao, G., Fleshner, P., Stappenbeck, T. S., McGovern, D. P. B., Keshavarzian, A., Mutlu, E. A., Sauk, J., Gevers, D., Xavier, R. J., Wang, D., Parkes, M., & Virgin, H. W. (2015). Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. *Cell*, *160*(3), 447–460. https://doi.org/10.1016/j.cell.2015.01.002

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, *27*(5), 824–834. https://doi.org/10.1101/gr.213959.116

Ofir, G., & Sorek, R. (2018). Contemporary Phage Biology: From Classic Models to New Insights. *Cell*, *172*(6), 1260–1270. https://doi.org/10.1016/j.cell.2017.10.045

Oksanen, H. M. & ICTV Report Consortium. (2017). ICTV Virus Taxonomy Profile: Corticoviridae. *Journal of General Virology*, *98*(5), 888–889. https://doi.org/10.1099/jgv.0.000795

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733-745. https://doi.org/10.1093/nar/gkv1189

Olszak, T., Latka, A., Roszniowski, B., Valvano, M. A., & Drulis-Kawa, Z. (2017). Phage Life Cycles Behind Bacterial Biodiversity. *Current Medicinal Chemistry*, *24*(36). https://doi.org/10.2174/0929867324666170413100136

Onallah, H., Hazan, R., Nir-Paz, R., Israeli Phage Therapy Center (IPTC) Study Team, Yerushalmy, O., Rimon, A., Braunstein, R., Gelman, D., Alkalay, S., Abdalrhman, M., Stuczynski, D., Coppenhagen-Glazer, S., Gelman, S., Khalifa, L., Adler, K., Yerushalmy, O., Rimon, A., Braunstein, R., Alkalay, S., … Livni, G. (2023). Compassionate Use of Bacteriophages for Failed Persistent Infections During the First 5 Years of the Israeli Phage Therapy Center. *Open Forum Infectious Diseases*, *10*(5), ofad221. https://doi.org/10.1093/ofid/ofad221

Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., & Phillippy, A. M. (2019). Mash Screen: High-throughput sequence containment estimation for genome discovery. *Genome Biology*, *20*(1), 232. https://doi.org/10.1186/s13059-019-1841-x

Oppenheim, A. B., Kobiler, O., Stavans, J., Court, D. L., & Adhya, S. (2005). Switches in Bacteriophage Lambda Development. *Annual Review of Genetics*, *39*(1), 409–429. https://doi.org/10.1146/annurev.genet.39.073003.113656

Park, J. Y., Moon, B. Y., Park, J. W., Thornton, J. A., Park, Y. H., & Seo, K. S. (2017). Genetic engineering of a temperate phage-based delivery system for CRISPR/Cas9 antimicrobials against Staphylococcus aureus. *Scientific Reports*, *7*(1), 44929. https://doi.org/10.1038/srep44929

Payaslian, F., Gradaschi, V., & Piuri, M. (2021). Genetic manipulation of phages for therapy using BRED. *Current Opinion in Biotechnology*, *68*, 8–14. https://doi.org/10.1016/j.copbio.2020.09.005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, *12*(null), 2825–2830.

Pereira Guedes, T., Alves Silva, J., Neves, S., Falcão, D., Costa, P., Lago, P., Pedroto, I., & Salgado, M. (2023). Positioning Aeromonas Infection in Inflammatory Bowel Disease: A

Retrospective Analysis. *GE - Portuguese Journal of Gastroenterology*, *30*(1), 20–28. https://doi.org/10.1159/000520272

Peschel, S., Müller, C. L., Von Mutius, E., Boulesteix, A.-L., & Depner, M. (2021). NetCoMi: Network construction and comparison for microbiome data in R. *Briefings in Bioinformatics*, *22*(4), bbaa290. https://doi.org/10.1093/bib/bbaa290

Ranveer, S. A., Dasriya, V., Ahmad, M. F., Dhillon, H. S., Samtiya, M., Shama, E., Anand, T., Dhewa, T., Chaudhary, V., Chaudhary, P., Behare, P., Ram, C., Puniya, D. V., Khedkar, G. D., Raposo, A., Han, H., & Puniya, A. K. (2024). Positive and negative aspects of bacteriophages and their immense role in the food chain. *Npj Science of Food*, *8*(1), 1. https://doi.org/10.1038/s41538-023-00245-8

Rebersek, M. (2021). Gut microbiome and its role in colorectal cancer. *BMC Cancer*, *21*(1), 1325. https://doi.org/10.1186/s12885-021-09054-2

Ren, J., Song, K., Deng, C., Ahlgren, N. A., Fuhrman, J. A., Li, Y., Xie, X., Poplin, R., & Sun, F. (2020). Identifying viruses from metagenomic data using deep learning. *Quantitative Biology*, *8*(1), 64–77. https://doi.org/10.1007/s40484-019-0187-4

Rish, I. (2001). *An empirical study of the naive Bayes classifier*. 6.

Roux, S., Adriaenssens, E. M., Dutilh, B. E., Koonin, E. V., Kropinski, A. M., Krupovic, M., Kuhn, J. H., Lavigne, R., Brister, J. R., Varsani, A., Amid, C., Aziz, R. K., Bordenstein, S. R., Bork, P., Breitbart, M., Cochrane, G. R., Daly, R. A., Desnues, C., Duhaime, M. B., … Eloe-Fadrosh, E. A. (2019). Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nature Biotechnology*, *37*(1), 29–37. https://doi.org/10.1038/nbt.4306

Roux, S., Camargo, A. P., Coutinho, F. H., Dabdoub, S. M., Dutilh, B. E., Nayfach, S., & Tritt, A. (2022). iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes. *bioRxiv*, 2022.07.28.501908. https://doi.org/10.1101/2022.07.28.501908

Roux, S., Enault, F., Hurwitz, B. L., & Sullivan, M. B. (2015). VirSorter: Mining viral signal from microbial genomic data. *PeerJ*, *3*, e985. https://doi.org/10.7717/peerj.985

Roux, S., Krupovic, M., Daly, R. A., Borges, A. L., Nayfach, S., Schulz, F., Sharrar, A., Matheus Carnevali, P. B., Cheng, J.-F., Ivanova, N. N., Bondy-Denomy, J., Wrighton, K. C., Woyke, T., Visel, A., Kyrpides, N. C., & Eloe-Fadrosh, E. A. (2019). Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nature Microbiology*, *4*(11), 1895–1906. https://doi.org/10.1038/s41564-019-0510-x

Roux, S., Krupovic, M., Poulet, A., Debroas, D., & Enault, F. (2012). Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PLoS ONE*, *7*(7), e40418. https://doi.org/10.1371/journal.pone.0040418

Rowan, F., Docherty, N. G., Murphy, M., Murphy, B., Coffey, J. C., & O'Connell, P. R. (2010). Desulfovibrio Bacterial Species Are Increased in Ulcerative Colitis. *Diseases of the Colon & Rectum*, *53*(11), 1530–1536. https://doi.org/10.1007/DCR.0b013e3181f1e620

Rozanov, D. V., D'Ari, R., & Sineoky, S. P. (1998). RecA-Independent Pathways of Lambdoid Prophage Induction in *Escherichia coli*. *Journal of Bacteriology*, *180*(23), 6306–6315. https://doi.org/10.1128/JB.180.23.6306-6315.1998

Russell, D. A., & Hatfull, G. F. (2017). PhagesDB: the actinobacteriophage database. *Bioinformatics (Oxford, England)*, *33*(5), 784–786. PubMed. https://doi.org/10.1093/bioinformatics/btw711

Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, *26*(2), 544–548. https://doi.org/10.1093/nar/26.2.544

Sanmukh, S., Khairnar, K., Paunikar, W., & Lokhande, S. (2015). Understanding carbon regulation in aquatic systems—Bacteriophages as a model. *F1000Research*, *4*, 138. https://doi.org/10.12688/f1000research.6031.1

Santiago-Rodriguez, T. M., & Hollister, E. B. (2022). Unraveling the viral dark matter through viral metagenomics. *Frontiers in Immunology*, *13*, 1005107. https://doi.org/10.3389/fimmu.2022.1005107

Sato, Y., Tsujinaka, S., Miura, T., Kitamura, Y., Suzuki, H., & Shibata, C. (2023). Inflammatory Bowel Disease and Colorectal Cancer: Epidemiology, Etiology, Surveillance, and Management. *Cancers*, *15*(16), 4154. https://doi.org/10.3390/cancers15164154

Sausset, R., Petit, M. A., Gaboriau-Routhiau, V., & De Paepe, M. (2020). New insights into intestinal phages. *Mucosal Immunology*, *13*(2), 205–215. https://doi.org/10.1038/s41385-019-0250-5

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/nar/gkab1112

Schmitt, M., & Greten, F. R. (2021). The inflammatory pathogenesis of colorectal cancer. *Nature Reviews Immunology*, *21*(10), 653–667. https://doi.org/10.1038/s41577-021-00534-x

Schubert, R. A., Dodd, I. B., Egan, J. B., & Shearwin, K. E. (2007). Cro's role in the CI–Cro bistable switch is critical for λ's transition from lysogeny to lytic development. *Genes & Development*, *21*(19), 2461–2472. https://doi.org/10.1101/gad.1584907

Secor, P. R., Burgener, E. B., Kinnersley, M., Jennings, L. K., Roman-Cruz, V., Popescu, M., Van Belleghem, J. D., Haddock, N., Copeland, C., Michaels, L. A., De Vries, C. R., Chen, Q., Pourtois, J., Wheeler, T. J., Milla, C. E., & Bollyky, P. L. (2020). Pf Bacteriophage and Their Impact on Pseudomonas Virulence, Mammalian Immunity, and Chronic Infections. *Frontiers in Immunology*, *11*, 244. https://doi.org/10.3389/fimmu.2020.00244

Shang, J., Tang, X., & Sun, Y. (2023). PhaTYP: predicting the lifestyle for bacteriophages using BERT. *Briefings in Bioinformatics*, *24*(1), bbac487. https://doi.org/10.1093/bib/bbac487

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498–2504. https://doi.org/10.1101/gr.1239303

Shao, Q., Trinh, J. T., McIntosh, C. S., Christenson, B., Balázsi, G., & Zeng, L. (2017). Lysis-lysogeny coexistence: Prophage integration during lytic development. *MicrobiologyOpen*, *6*(1). https://doi.org/10.1002/mbo3.395

Shao, Q., Trinh, J. T., & Zeng, L. (2019). High-resolution studies of lysis-lysogeny decision-making in bacteriophage lambda. *The Journal of Biological Chemistry*, *294*(10), 3343–3349. PubMed. https://doi.org/10.1074/jbc.TM118.003209

Shkoporov, A. N., & Hill, C. (2019). Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host & Microbe*, *25*(2), 195–209. https://doi.org/10.1016/j.chom.2019.01.017

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., & Jemal, A. (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, *70*(3), 145–164. https://doi.org/10.3322/caac.21601

Siegel, R. L., Wagle, N. S., Cercek, A., Smith, R. A., & Jemal, A. (2023). Colorectal cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, *73*(3), 233–254. https://doi.org/10.3322/caac.21772

Sieiro, C., Areal-Hermida, L., Pichardo-Gallardo, Á., Almuiña-González, R., De Miguel, T., Sánchez, S., Sánchez-Pérez, Á., & Villa, T. G. (2020). A Hundred Years of Bacteriophages: Can Phages Replace Antibiotics in Agriculture and Aquaculture? *Antibiotics*, *9*(8), 493. https://doi.org/10.3390/antibiotics9080493

Silveira, C. B., & Rohwer, F. L. (2016). Piggyback-the-Winner in host-associated microbial communities. *Npj Biofilms and Microbiomes*, *2*(1), 16010. https://doi.org/10.1038/npjbiofilms.2016.10

Smith, M. J., & Jeddeloh, J. A. (2005). DNA Methylation in Lysogens of Pathogenic *Burkholderia* spp. Requires Prophage Induction and Is Restricted to Excised Phage DNA. *Journal of Bacteriology*, *187*(3), 1196–1200. https://doi.org/10.1128/JB.187.3.1196-1200.2005

Smits, S. L., Bodewes, R., Ruiz-Gonzalez, A., BaumgÃ¤rtner, W., Koopmans, M. P., Osterhaus, A. D. M. E., & SchÃ¼rch, A. C. (2014). Assembly of viral genomes from metagenomes. *Frontiers in Microbiology*, *5*. https://doi.org/10.3389/fmicb.2014.00714

Soliman, W. S., Shaapan, R. M., Mohamed, L. A., & Gayed, S. S. R. (2019). Recent biocontrol measures for fish bacterial diseases, in particular to probiotics, bio-encapsulated vaccines, and phage therapy. *Open Veterinary Journal*, *9*(3), 190. https://doi.org/10.4314/ovj.v9i3.2

Song, K. (2020). Classifying the Lifestyle of Metagenomically-Derived Phages Sequences Using Alignment-Free Methods. *Frontiers in Microbiology*, *11*, 567769. https://doi.org/10.3389/fmicb.2020.567769

Stanton, C. R., Batinovic, S., & Petrovski, S. (2023). Burkholderia contaminans Bacteriophage CSP3 Requires O-Antigen Polysaccharides for Infection. *Microbiology Spectrum*, *11*(3), e05332-22. https://doi.org/10.1128/spectrum.05332-22

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028. https://doi.org/10.1038/nbt.3988

Strathdee, S. A., Hatfull, G. F., Mutalik, V. K., & Schooley, R. T. (2023). Phage therapy: From biological mechanisms to future directions. *Cell*, *186*(1), 17–31. https://doi.org/10.1016/j.cell.2022.11.017

Summers, W. C. (2012). The strange history of phage therapy. *Bacteriophage*, *2*(2), 130–133. https://doi.org/10.4161/bact.20757

Swenson, K., Gonzalez, J.-P., & Steen, T. Y. (2024). A Way Forward for Phage Therapy in the United States. *Georgetown Medical Review*, *8*(1). https://doi.org/10.52504/001c.117696

Tang, X., Zhong, L., Tang, L., Fan, C., Zhang, B., Wang, M., Dong, H., Zhou, C., Rensing, C., Zhou, S., & Zeng, G. (2023). Lysogenic bacteriophages encoding arsenic resistance determinants promote bacterial community adaptation to arsenic toxicity. *The ISME Journal*, *17*(7), 1104–1115. https://doi.org/10.1038/s41396-023-01425-w

Terzian, P., Olo Ndela, E., Galiez, C., Lossouarn, J., Pérez Bucio, R. E., Mom, R., Toussaint, A., Petit, M.-A., & Enault, F. (2021). PHROG: families of prokaryotic virus proteins clustered using remote homology. *NAR Genomics and Bioinformatics*, *3*(3), lqab067. https://doi.org/10.1093/nargab/lqab067

Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., & Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications*, *13*(1), 2561. https://doi.org/10.1038/s41467-022-30269-9

Tétart, F., Repoila, F., Monod, C., & Krisch, H. M. (1996). Bacteriophage T4 Host Range is Expanded by Duplications of a Small Domain of the Tail Fiber Adhesin. *Journal of Molecular Biology*, *258*(5), 726–731. https://doi.org/10.1006/jmbi.1996.0281

Tetz, G., Brown, S. M., Hao, Y., & Tetz, V. (2018). Parkinson's disease and bacteriophages as its overlooked contributors. *Scientific Reports*, *8*(1), 10812. https://doi.org/10.1038/s41598-018-29173-4

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., … Zhang, J. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Thompson, L. R., Zeng, Q., Kelly, L., Huang, K. H., Singer, A. U., Stubbe, J., & Chisholm, S. W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences*, *108*(39), E757–E764. https://doi.org/10.1073/pnas.1102164108

Trgovec-Greif, L., Hellinger, H.-J., Mainguy, J., Pfundner, A., Frishman, D., Kiening, M., Webster, N. S., Laffy, P. W., Feichtinger, M., & Rattei, T. (2024). VOGDB—Database of Virus Orthologous Groups. *Viruses*, *16*(8), 1191. https://doi.org/10.3390/v16081191

Turner, D., Shkoporov, A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfenas-Zerbini, P., Van Zyl, L. J., Aziz, R. K., Oksanen, H. M., Poranen, M. M., Kropinski, A. M., Barylski, J., Brister, J. R., Chanisvili, N., Edwards, R. A., Enault, F., Gillis, A., Knezevic, P., Krupovic, M., … Adriaenssens, E. M. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Archives of Virology*, *168*(2), 74. https://doi.org/10.1007/s00705-022-05694-2

Twort, F. W. (1961). An investigation on the nature of ultra-microscopic viruses. *Acta Kravsi*.

Tynecki, P., Guziński, A., Kazimierczak, J., Jadczuk, M., Dastych, J., & Onisko, A. (2020). *PhageAI - Bacteriophage Life Cycle Recognition with Machine Learning and Natural Language Processing* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.07.11.198606

Unterer, M., Khan Mirzaei, M., & Deng, L. (2023). Targeted Single-Phage Isolation Reveals Phage-Dependent Heterogeneous Infection Dynamics. *Microbiology Spectrum*, *11*(3), e05149-22. https://doi.org/10.1128/spectrum.05149-22

US Food and Drug Administration. (2019). *Adaptive Designs for Clinical Trials of Drugs and Biologics: Guidance for Industry.* https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry

Versoza, C. J., & Pfeifer, S. P. (2022). Computational Prediction of Bacteriophage Host Ranges. *Microorganisms*, *10*(1), 149. https://doi.org/10.3390/microorganisms10010149

Villarroel, J., Kleinheinz, K., Jurtz, V., Zschach, H., Lund, O., Nielsen, M., & Larsen, M. (2016). HostPhinder: A Phage Host Prediction Tool. *Viruses*, *8*(5), 116. https://doi.org/10.3390/v8050116

Villarroel, J., Larsen, M. V., Kilstrup, M., & Nielsen, M. (2017). Metagenomic Analysis of Therapeutic PYO Phage Cocktails from 1997 to 2014. *Viruses*, *9*(11), E328. https://doi.org/10.3390/v9110328

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., Van Der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waldbauer, J. R., Coleman, M. L., Rizzo, A. I., Campbell, K. L., Lotus, J., & Zhang, L. (2019). Nitrogen sourcing during viral infection of marine cyanobacteria. *Proceedings of the National Academy of Sciences*, *116*(31), 15590–15595. https://doi.org/10.1073/pnas.1901856116

Wang, W., Chen, L., Zhou, R., Wang, X., Song, L., Huang, S., Wang, G., & Xia, B. (2014). Increased Proportions of Bifidobacterium and the Lactobacillus Group and Loss of Butyrate-Producing Bacteria in Inflammatory Bowel Disease. *Journal of Clinical Microbiology*, *52*(2), 398–406. https://doi.org/10.1128/JCM.01500-13

Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Weinbauer, M. G. (2004). Ecology of prokaryotic viruses. *FEMS Microbiology Reviews*, *28*(2), 127–181. https://doi.org/10.1016/j.femsre.2003.08.001

Williams, K. P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: Sublocation preference of integrase subfamilies. *Nucleic Acids Research*, *30*(4), 866–875. https://doi.org/10.1093/nar/30.4.866

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

World Health Organization. (2023, November 21). *Antimicrobial resistance*. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

World Health Organization. (2024, April 26). *WHO reports widespread overuse of antibiotics in patients hospitalized with COVID-19*. https://www.who.int/news/item/26-04-2024-who-reports-widespread-overuse-of-antibiotics-in-patients--hospitalized-with-covid-19

Wu, S., Fang, Z., Tan, J., Li, M., Wang, C., Guo, Q., Xu, C., Jiang, X., & Zhu, H. (2021). DeePhage: Distinguishing virulent and temperate phage-derived sequences in metavirome data with a deep learning approach. *GigaScience*, *10*(9), giab056. https://doi.org/10.1093/gigascience/giab056

Yang, D., Wang, S., Sun, E., Chen, Y., Hua, L., Wang, X., Zhou, R., Chen, H., Peng, Z., & Wu, B. (2022). A temperate *Siphoviridae* bacteriophage isolate from Siberian tiger enhances the virulence of methicillin-resistant *Staphylococcus aureus* through distinct mechanisms. *Virulence*, *13*(1), 137–148. https://doi.org/10.1080/21505594.2021.2022276

Yang, M., Derbyshire, M. K., Yamashita, R. A., & Marchler-Bauer, A. (2020). NCBI's Conserved Domain Database and Tools for Protein Domain Analysis. *Current Protocols in Bioinformatics*, *69*(1), e90. https://doi.org/10.1002/cpbi.90

Yuan, Y., Peng, Q., Wu, D., Kou, Z., Wu, Y., Liu, P., & Gao, M. (2015). Effects of Actin-Like Proteins Encoded by Two Bacillus pumilus Phages on Unstable Lysogeny, Revealed by Genomic Analysis. *Applied and Environmental Microbiology*, *81*(1), 339–350. https://doi.org/10.1128/AEM.02889-14

Yukgehnaish, K., Rajandas, H., Parimannan, S., Manickam, R., Marimuthu, K., Petersen, B., Clokie, M. R. J., Millard, A., & Sicheritz-Pontén, T. (2022). PhageLeads: Rapid Assessment of Phage Therapeutic Suitability Using an Ensemble Machine Learning Approach. *Viruses*, *14*(2). https://doi.org/10.3390/v14020342

Żaczek, M., Weber-Dąbrowska, B., Międzybrodzki, R., Łusiak-Szelachowska, M., & Górski, A. (2020). Phage Therapy in Poland – a Centennial Journey to the First Ethically Approved Treatment Facility in Europe. *Frontiers in Microbiology*, *11*, 1056. https://doi.org/10.3389/fmicb.2020.01056

Zajmi, A., Teo, J., & Yeo, C. C. (2022). Epidemiology and Characteristics of Elizabethkingia spp. Infections in Southeast Asia. *Microorganisms*, *10*(5), 882. https://doi.org/10.3390/microorganisms10050882

Zaldastanishvili, E., Leshkasheli, L., Dadiani, M., Nadareishvili, L., Askilashvili, L., Kvatadze, N., Goderdzishvili, M., Kutateladze, M., & Balarjishvili, N. (2021). Phage Therapy Experience at the Eliava Phage Therapy Center: Three Cases of Bacterial Persistence. *Viruses*, *13*(10), 1901. https://doi.org/10.3390/v13101901

Zhao, G., Vatanen, T., Droit, L., Park, A., Kostic, A. D., Poon, T. W., Vlamakis, H., Siljander, H., Härkönen, T., Hämäläinen, A.-M., Peet, A., Tillmann, V., Ilonen, J., Wang, D., Knip, M., Xavier, R. J., & Virgin, H. W. (2017). Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proceedings of the National Academy of Sciences*, *114*(30). https://doi.org/10.1073/pnas.1706359114

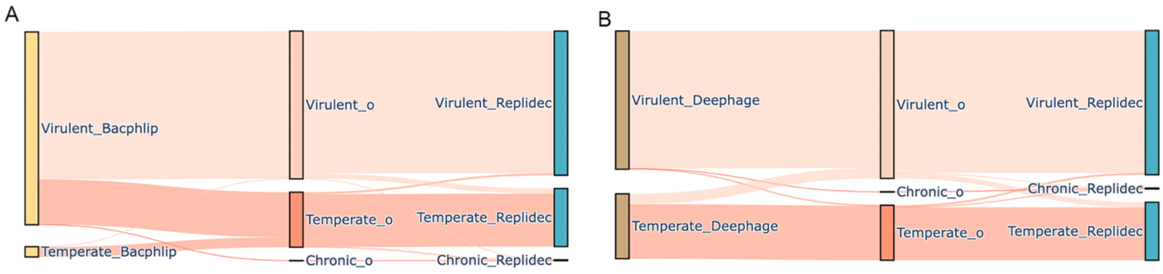
Zheng, J., Sun, Q., Zhang, M., Liu, C., Su, Q., Zhang, L., Xu, Z., Lu, W., Ching, J., Tang, W., Cheung, C. P., Hamilton, A. L., Wilson O'Brien, A. L., Wei, S. C., Bernstein, C. N., Rubin, D. T., Chang, E. B., Morrison, M., Kamm, M. A., … Ng, S. C. (2024). Noninvasive, microbiome-based diagnosis of inflammatory bowel disease. *Nature Medicine*. https://doi.org/10.1038/s41591-024-03280-4

Zhou, F., Gan, R., Zhang, F., Ren, C., Yu, L., Si, Y., & Zhiwei, H. (n.d.). *PHISDetector: A tool to detect diverse in silico phage-host interaction signals for virome studies*. 17.

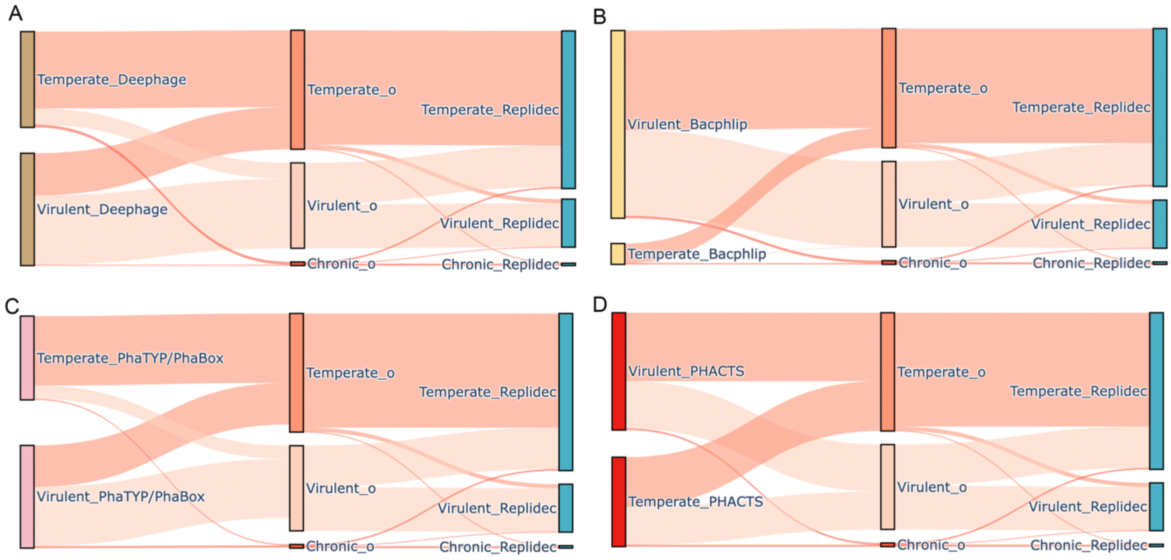
Zschach, H., Joensen, K. G., Lindhard, B., Lund, O., Goderdzishvili, M., Chkonia, I., Jgenti, G., Kvatadze, N., Alavidze, Z., Kutter, E. M., Hasman, H., & Larsen, M. V. (2015). What Can We Learn from a Metagenomic Analysis of a Georgian Bacteriophage Cocktail? *Viruses*, *7*(12), 6570–6589. https://doi.org/10.3390/v7122958

Zuo, T., Wong, S. H., Lam, K., Lui, R., Cheung, K., Tang, W., Ching, J. Y. L., Chan, P. K. S., Chan, M. C. W., Wu, J. C. Y., Chan, F. K. L., Yu, J., Sung, J. J. Y., & Ng, S. C. (2017). Bacteriophage transfer during faecal microbiota transplantation in *Clostridium difficile* infection is associated with treatment outcome. *Gut*, gutjnl-2017-313952. https://doi.org/10.1136/gutjnl-2017-313952

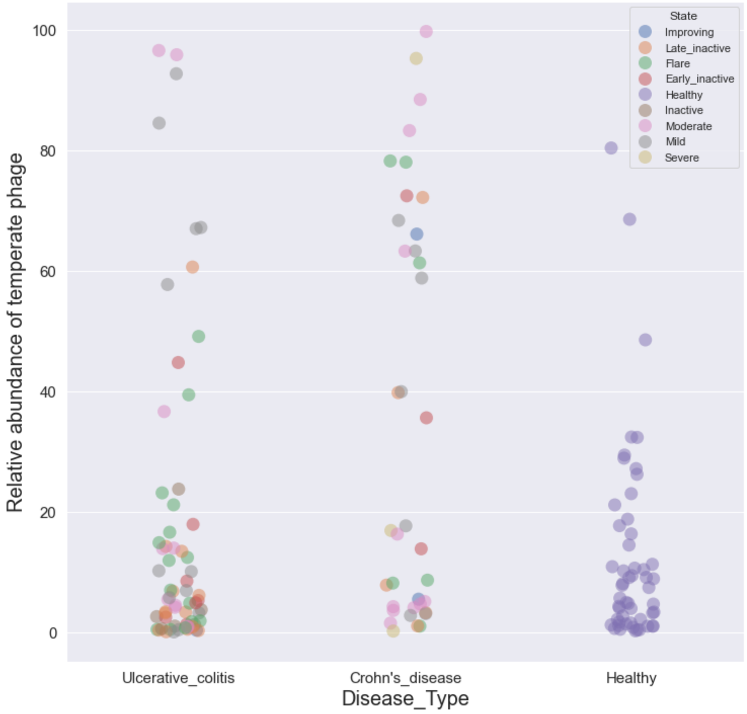
Appendix A: Figures



Appendix Figure 1 **Sankey plots comparison in test Dataset 1.**
Sankey plots show the comparison between BACHLIP(A, left) or Deephage(B, left), RepliDec (A and B, right), and benchmark label (A and B, middle with o as suffix) of simulate metagenomic contigs generated from test Dataset1.

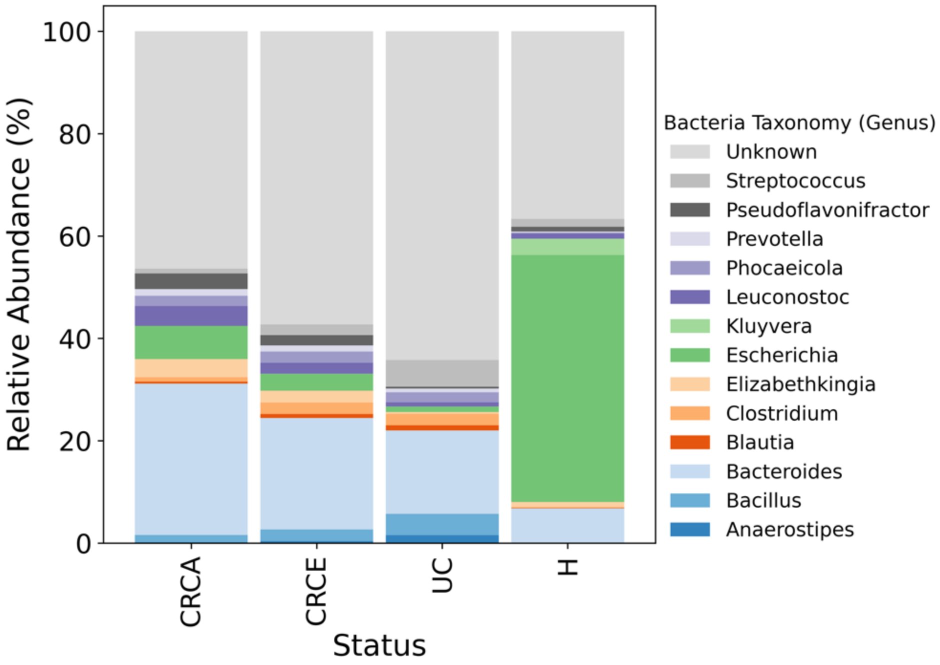


Appendix Figure 2 **Sankey plots comparison in test Dataset 2.**
Sankey plots compare different tools with benchmark labels (ABCD, middle with o as suffix) of simulated metagenomic contigs generated from test Dataset 2. DeePhage (A, left), BACHLIP (B, left), PhaTYP/PhaBox (B, left), and PHACTS (D, left) were on the left side, and RepliDec (all, right) is on the right column on all four subplots.



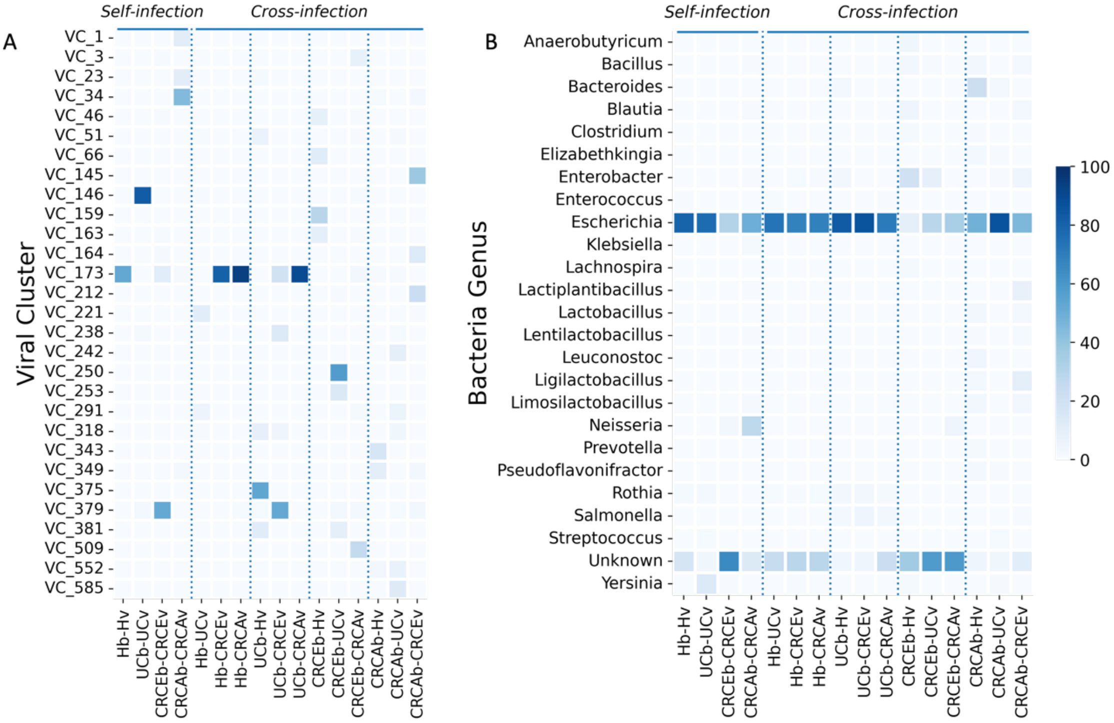
Appendix Figure 3 **Abundance profile of temperate replication cycles in both the IBD and healthy cohorts.**

The color indicates different health statuses.



Appendix Figure 4 **Bacteria composition in CRCA, CRCE, UC, and H.**

The relative abundance of bacteria at the genus level in metagenomic fecal samples. "CRCA" stands for advanced CRC status, "CRCE" also represents CRC in an early stage, "UC" refers to ulcerative colitis, and "H" for Healthy controls.



Appendix Figure 5 **Phage and bacteria composition in VT cross-infection samples.** The heatmap shows the high relative abundance of phages at the strain level (98% ANI) (A) and bacteria (B) in VT samples. The color indicates the relative abundance. The sample names in VT indicate the bacteria and viral source using "b" and "v" as suffixes. "CRCA" stands for advanced CRC status, "CRCE" also represents CRC in an early stage, "UC" refers to ulcerative colitis, and "H" for Healthy controls.

Appendix B: Tables

Because the table is too large to put in the Word file, the large tables are uploaded to GitHub: https://github.com/pengSherryYel/Data-for-Thesis. All files are under the Table folder.

Appendix Table 1 Bacteriophage accession IDs in RefSeq used in the training dataset.

Appendix Table 2 Bacteria and archaea accession IDs in RefSeq used in the training dataset.

Appendix Table 3 Test datasets accession IDs. Test dataset 1 (testD1) contains 470 RefSeq viral genomes from Mavrich whose replication cycle was experimented validated and consistent with the bioinformatic prediction results; Test dataset 2 (testD2) contains 610 represented viral genomes (can present 2920 genomes), which shows at most 95% similarity between each other also not similar to RefSeq and RepliDec training dataset.

Appendix Table 4 Sample accession in IBD study.

Appendix Table 5 Protein annotation of training datasets.

Appendix Table 6 Temperate phages predicted from commercial phage cocktails.

Acknowledgments

The completion of my thesis marks the end of my 3.5-year doctoral journey. As I reflect on my time in Germany, I feel deeply grateful. I want to sincerely thank everyone who supported and encouraged me throughout this process. Their invaluable guidance and assistance made this thesis possible.

I would like to sincerely thank my supervisor, Prof. Dr. Li Deng, for the wonderful opportunity to study in her group and participate in the ITN VIROINF program. I would like to express my heartfelt gratitude to my second supervisors, Dr. Jinlong Ru and Dr. Mohammadali Khan Mirzaei, for their invaluable support throughout my doctoral journey. During my first year, Dr. Jinlong Ru guided me, helping me adapt to the lab and life in Germany with patience. Following that, Dr. Mohammadali Khan Mirzaei took over as my supervisor, guiding me through the final stages of my studies. His expertise and constructive feedback were crucial in shaping my research direction. I am very grateful that he always guides me on the project and is willing to spend time discussing the projects.

I am also grateful to PD Dr. Jürgen Lassak, who kindly agreed to be the first examiner of the examination committee for my doctoral dissertation. I am grateful to PD Dr. Jürgen Lassak and Prof. Dr. Caroline Friedel for being such supportive members of the TAC meeting during my entire PhD journey. Their valuable suggestions have truly helped steer my studies in the right direction. I truly appreciate the generosity of Prof. Dr. Silke Robatzek, Prof. Dr. John Parsch, Prof. Dr. Jörg Nickelsen, and Prof. Dr. Dirk Metzler, who have graciously agreed to be the examiners for my doctoral dissertation.

I would like to express my appreciation to my colleagues Dr. Jinling Xue, Shiqi Luo, Wanqi Huang, Magdalena Unterer, Kawtar Tiamani, Adrian Thaqi, and Sophie Smith for their collaboration, particularly in the laboratory work. I would also like to thank Hulya Ali for her assistance with all the documentation and for providing additional administrative support.

I am also deeply grateful to Prof. Dr. Manja Marz, the coordinator of the VIROINF program, and Dr. Winfried Goettsch, who is the network manager of the VIROINF program. I also want to thank all the supervisors involved in the VIROINF program. They organize numerous wonderful workshops that provide me with an excellent opportunity to receive training as an early-stage

scientist. These workshops cover various topics essential to my development in the field, including research methodologies, data analysis, and collaboration techniques. Participating in these sessions enhances my skills and knowledge and connects me with experienced professionals and other aspiring scientists.

I am grateful to Prof. Dr. David L. Robertson. He gave me a great opportunity to study at the Centre of Virus Research (CVR) at the University of Glasgow. During my stay, I learned many techniques used in deep learning. I would also like to thank my colleagues at CVR, Dan Liu, Kieran Lamb, and Francesca Young, for supporting my secondment stay at the University of Glasgow.

I also appreciate the assistance and resources provided by Helmholtz Zentrum München and Ludwig-Maximilians-Universität, which greatly facilitated my work and enhanced my professional skills.

I have been incredibly blessed with supportive friends and family who have always encouraged me to pursue my goals in life. I am deeply grateful to my partner, Hanxiong Huang, for consistently being my rock and for providing unwavering support and encouragement. From the bottom of my heart, I would like to thank my parents for their love and support. I believe my father would be very happy to see this thesis completed if he could.

List of publications

Peng, X., Ru, J., Mirzaei, M. K., & Deng, L. (2022). RepliDec-Use naive Bayes classifier to identify virus lifecycle from metagenomics data. *bioRxiv*, 2022-07.

Peng, X., Smith, S. E., Huang, W., Ru, J., Mirzaei, M. K., & Deng, L. (2024). Metagenomic analyses of single phages and phage cocktails show instances of contamination with temperate phages and bacterial DNA. *bioRxiv*, 2024-09.

Luo, S., Ru, J., Mirzaei, M. K., Xue, J., **Peng, X**., Ralser, A., ... & Deng, L. (2023). Gut virome profiling identifies an association between temperate phages and colorectal cancer promoted by Helicobacter pylori infection. *Gut Microbes*, *15*(2), 2257291.