Technische Universität München
TUM School of Life Sciences

Exploring Phage-Bacteria Interactions: Insights Into Unculturable *Helicobacter* Phages and the Role of Phages in Gastrointestinal Diseases

Magdalena Unterer

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades einer

Doktorin der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitz:	Prof. Dr. Friederike Ebner
Prüfende der Dissertation:	1. Prof. Dr. Li Deng
	2. Prof. Dr. Markus Gerhard
	3. Prof. Dr. Julia Frunzke

Die Dissertation wurde am 23.10.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 05.02.2025 angenommen.

Dedicated to my parents

Acknowledgment

First and foremost, I would like to express my sincere gratitude to Prof. Dr. Li Deng, who gave me the opportunity to do my PhD in her group and made my PhD journey even possible. I am grateful for her supervision and unlimited support.

I also want to thank Dr. Mohammadali Khan Mirzaei for his supervision and for his constant support. Thank him for his patience when we had endless discussions about my work in the hallway or in our weekly meetings, his positive mindset to keep me pushing even in tough times, his honest feedback about my writing, and his faith in me that I can do it. Thank him for making me a better scientist, a better writer and helping me grow as a person. Thank you, Ali!

I also want to thank Prof. Dr. Markus Gerhard, Prof. Dr Benjamin Schusser, and Prof. Dr. Dietmar Zehn for being on my thesis committee and for their constructive feedback and engaging questions.

I am also grateful for all my friends, colleagues and supporting members of the group and institute. Thank you, Sophie, Kawtar, Adrian, and Sonja, for all the memories we made together, for all the fun times we had, for making stressful days still enjoyable, and for being by my side through this journey. A big shout out also to Sarah, Silke, and Monique for supporting me in every possible way; without you, *Helicobacter pylori* would not have flourished as they did. Thank you for taking up this challenge with me and having the patience to deal with my perfectionism.

Last but not least, I want to thank Thomas and my family. I am eternally grateful for their unlimited support, faith and resilience to walk this educational journey with me. Thank you for all the motivational speeches in tough times, for keeping the energy when I couldn’t, for listening to all of my over-scientifically filled stories, and for your scientific contributions to help me complete my projects. Thank you for having my back – always!

Abstract

Antimicrobial resistance (AMR) is the phenomenon where bacteria expand their defense reservoirs against current antibiotics. These resistances pose a major health threat and are predicted to cause 10 million deaths by 2050. Despite rising AMRs, new antibiotics rarely enter the market as the developmental process is complex and slow. Consequently, the need for alternative treatment options is more critical than ever.

An ancient alternative for bacterial inhibition, long before antibiotics were available, was bacteriophages (in short, phages). Phages are highly abundant viruses that can be found in the presence of bacteria. As their natural enemy, phages have successfully proven to clear bacterial infections. For phage therapy, phages need to be isolated against the pathogen which depicts one of the major challenges in phage therapy. Phage isolation is a tedious and time-consuming process that is reaching its current limits.

For that, I developed a technique called "targeted single phage isolation," where we can quickly isolate and characterize phages for various pathogens. We used natural sources such as fecal samples or sewage water, which we filtered, stained, and mixed with our target bacteria. Mixed cells were then separated by a flow cytometer into tagged cells (bacterial cells infected by phage) and non-tagged cells (bacterial cells only), a process called viral tagging. Viral-tagged cells were then studied individually. Via growth monitoring, infection dynamics were determined to identify phages that were actively infecting the bacteria. Those were then tested in the subsequent quantitative polymerase chain reaction (qPCR), where we screened for viral traits. These characterizations enabled us to quickly select potential candidates for phage therapy. Targeted single phage isolation also allowed us to observe phage individuality and revealed heterogenous infection dynamics, a behavioral difference between single-cell infection and group infection. This was observed for T1, T4, and T7 phages combined with *Escherichia coli* 11303.

We also used the developed viral tagging technique to detect phages for *Helicobacter pylori*. 491 *Helicobacter pylori* contigs were discovered. We compared all

contigs to the publicly available phage and prophage sequences and exposed a major discrepancy between them. Viral-tagged contigs shared little to no sequence similarities to the public sequences despite being annotated with phage attributes or having *H. pylori* as host prediction. We characterized viral-tagged contigs and public sequences and identified NAD-dependent epimerase/dehydratase as this dataset's only shared auxiliary metabolic gene (AMG). The remaining AMGs were all unique for each group. The dataset was also scanned for endolysins, an essential enzyme for bacterial cell wall degradation. 20 endolysins were found in total, whereas 19 belonged to the viral tagging dataset. These opposing views highlight how diverse phages can be and how important it is to have an extensive database to fully understand the role of phages.

Phages play a crucial part, especially in diseases. In my last project, we examined phages' role in gut microbiota and diseases. Bacterial cells and viral particles were isolated from colorectal cancer patients in early and late stages, Ulcerative Colitis patients, and healthy individuals. We sequenced the bacteriome and the virome and discovered a distinct picture for each condition. In addition, we self-infected and cross-infected bacteria with phages to investigate bacterial and viral abundance. *Escherichia* was detected to be the most abundant taxa in all samples, but no common viral cluster was identified. The subsequent interaction analysis revealed the microbial association network, which was unique for each condition. These findings suggest that diseases highly influence microbial compositions and connections.

Overall, the developed technique enabled us to expand our research. We could analyze phages on a single-cell level, extend phage reference databases, and look into the role of phages in diseases. We believe this is a solid foundation for advancing phage research, paving the way for revolutionary developments in phage therapy.

Zusammenfassung

Antibiotikaresistenzen nehmen rasant zu, was zur Folge haben wird, dass in 20 Jahren mehr als 10 Millionen Menschen jährlich daran sterben werden. Der Grund dafür ist, dass Bakterien zunehmend gegen immer mehr Antibiotika resistent werden, oft gegen mehrere gleichzeitig. Obwohl die Resistenzen ansteigen, kommen nur sehr wenige neue Antibiotika auf dem Markt, da die Prozesse zur Neuentwicklung von Antibiotika komplex und sehr langsam sind. Daher müssen alternative Behandlungsmethoden gesucht und erforscht werden.

Eine Alternative ist die Phagentherapie, die bereits erfolgreich angewendet wurde, lange bevor Antibiotika vorhanden waren. In dieser Behandlungsmethode werden Viren, sogenannte Phagen, gegen bakterielle Infektionen eingesetzt. Diese Viren sind die natürlichen Feinde von Bakterien und existieren überall dort, wo auch Bakterien vorkommen. Trotz ihrer weiten Verbreitung und erfolgreichen Hilfe gegen bakterielle Infektionen, ist die Isolierung und Identifizierung von Phagen sehr aufwendig und schwierig. Die aktuell verfügbaren Methoden sind nur begrenzt wirksam und stoßen zunehmend an ihre Grenzen.

Daher habe ich eine Methode entwickelt „gezielte Isolierung von einzelnen Phagen“. Durch diese Methode sind wir in der Lage, Phagen für viele verschiedene Bakterien zu isolieren und zu charakterisieren. Dafür verwendeten wir Abwasserproben sowie Stuhlproben, konzentrierten die vorhandenen Phagen, färbten sie ein und mischten sie mit den Bakterien unsrer Wahl. Mit einem Durchflusszytometer separierten wir Zellen, die mit einer Phage infiziert waren von jenen ohne Phage.

Durch die entwickelte Methode konnten wir das Verhalten von einzelnen Phagen während des Infektionszyklus beobachten und feststellen, dass Phagen heterogene Infektionsdynamiken besitzen. Des Weiteren konnten wir alle getesteten Phagen im weiteren Prozess charakterisieren und identifizieren.

Die entwickelte Methode wurde auch dazu verwendet Phagen für *Helicobacter pylori* zu identifizieren. Wir fanden 491 Contigs. Diese haben wir mit publizierten Sequenzen verglichen. Dabei trat eine Diskrepanz zwischen den von uns identifizierten Contigs und den bereits publizierten Sequenzen zutage. NAD-abhängige Epimerase/Dehydratase

wurde als einziges Gen in beiden Datensätzen gefunden, die restlichen identifizierten AMGs waren alle einzigartig. Zudem fanden wir erhebliche Differenzen in der Präsenz von endolytischen Enzymen, einem Protein für den bakteriellen Zellwandabbau. Von den 20 gefundenen Enzymen sind 19 den viral tagging Contigs zuzuordnen. Diese gegensätzlichen Resultate unterstreichen die immense Vielfalt der Phagen und zeigen wie wichtig es ist, Datenbanken permanent auszubauen.

Im letzten Projekt nutzten wir die entwickelte Methode, um Phagen und ihre Rolle in der Darmflora zu untersuchen. Dazu haben wir Stuhlproben von Patienten mit Darmkrebs im Früh- und Endstadium bearbeitet, sowie jene von Patienten mit ulzerative Kolitis und gesunden Menschen. Wir haben jeweils die Bakterien sowie Viren daraus isoliert und diese miteinander kreuzinfiziert, um die Interaktionen zwischen Wirt und Phagen in den jeweiligen Bedingungen zu analysieren. Die Daten zeigten, dass die bakterielle und virale Zusammensetzung je nach Erkrankung variiert und einzigartige Interaktionsnetzwerke bestehen. Diese Ergebnisse deuten darauf hin, dass die mikrobielle Komposition von Krankheiten beeinflusst wird.

Zusammenfassend haben wir eine Technik entwickelt, die uns tiefere Einblicke in Phagen und ihre Eigenschaften ermöglicht und auch zukünftig dabei helfen wird, Zusammenhänge besser zu erforschen. Wir sind zuversichtlich, dass diese Technologie für die Phagentherapie hilfreich sein wird.

Scientific Contributions

Magdalena Unterer, Mohammadali Khan Mirzaei, Li Deng
Targeted Single-Phage Isolation Reveals Phage-Dependent Heterogeneous Infection Dynamics
Microbiology Spectrum, DOI: https://doi.org/10.1128/spectrum.05149-22, (April 17th 2023)

Sophie E Smith, Wanqi Huang, Kawtar Tiamani, **Magdalena Unterer**, Mohammadali Khan Mirzaei, Li Deng
Emerging technologies in the study of the virome
Current Opinion in Virology, DOI https://doi.org/10.1016/j.coviro.2022.101231, (May 25th 2022)

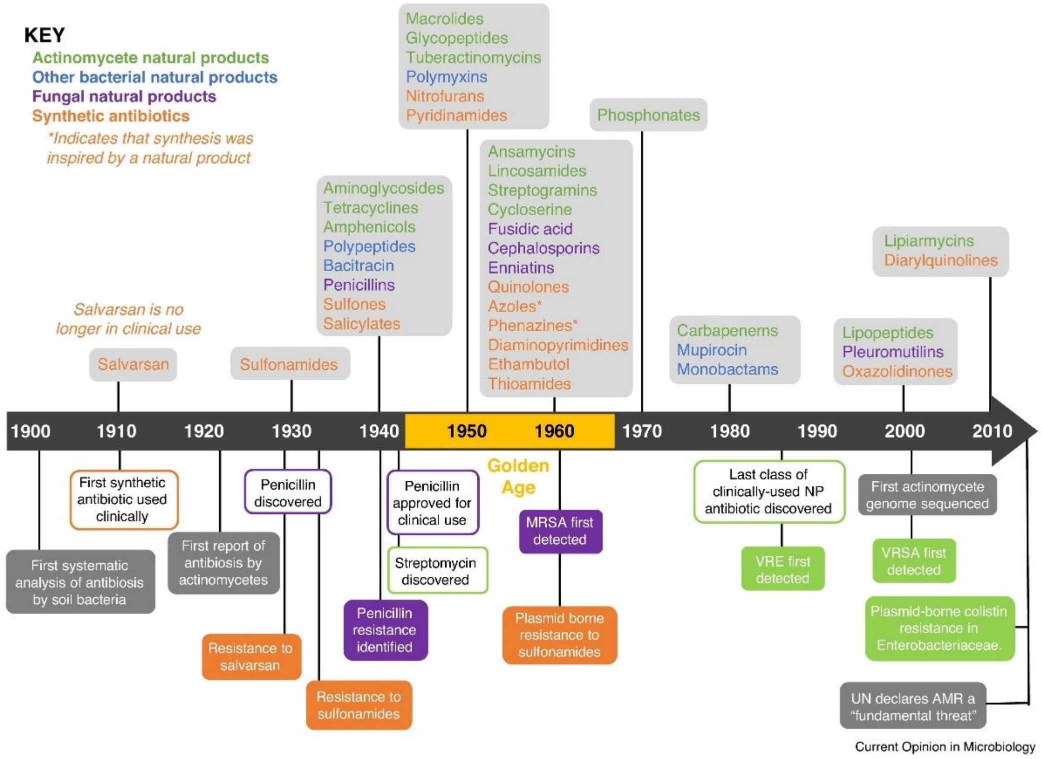
Magdalena Unterer, Mohammadali Khan Mirzaei and Li Deng
Gut Phage Database: phage mining in the cave of wonders
Signal Transduction and Targeted Therapy, DOI https://doi.org/10.1038/s41392-021-00615-2, (May 17th 2021)

Table of Contents

1. Introduction

Are bacterial infections life-threatening again? For most of human existence, bacterial infections were often deadly. Millions died due to pneumonia or infected wounds, especially during war times, until penicillin was found. Penicillin was discovered in 1928 as the first of its kind [1]. Penicillin is a molecule produced by a fungus and capable of inhibiting the growth of bacteria. It was the first antibiotic (antimicrobial substance) that saved thousands of lives, built the base for many other antibiotics, and knocked off the start of the golden era of antibiotics [1]. For decades, bacterial infections were treated easily and became innocuous. An entire palette of antibiotics was available with various modes of action – stopping the cell wall construction, interfering with deoxyribonucleic acid (DNA) replication, or interrupting protein synthesis. However, this growing palette had an alarming reason - antibiotic resistance [1], [2]. Bacteria developed strategies to escape all kinds of antibiotic treatments. Consequently, new drugs were designed to interfere with the latest defense mechanisms developed by bacteria. Nevertheless, as bacteria constantly evolve, their pool of defense actions has grown as well, but the pool of drugs has not, as pictured in Figure 1 [3]. That leaves us with multi-resistant bacteria, which are regaining their status as a death threat.

Antibiotic-resistant bacteria are detected everywhere on the globe [4], [5] caused by the massive overuse and misuse of antibiotics in hospitals or livestock [6], [7]. Every day, more and more join the group of multidrug-resistant pathogens [8]. This growing pool of resistant pathogens is an issue that is becoming more and more pressing [3]. Almost five million deaths were triggered by resistant pathogens in 2019 globally. Out of those five million, 1.2 million were directly caused by resistant strains. Multidrug-resistant bacteria kill more people globally each year than HIV/Aids (Human Immunodeficiency Virus -864,000 deaths) or Malaria (643,000 deaths) [6], [9]. In Germany, 30,000 – 35,000 people are diagnosed with an infection caused by multidrug-resistant bacteria each year [10]. Roughly a third dies because of that infection. The predictions are that without counter-measurements, those pathogens will kill up to 10 million by 2050 worldwide [5], [6], [9]. Multi-resistant pathogens have put the world in a difficult position as these infections burden each country's health system and economic status [7].



Current Opinion in Microbiology

Figure 1_Antibiotic Resistance Timeline

The timeline shows the discovery of new antibiotics. The different colors represent the source for the antibiotic: green = actinomycetes, blue = other bacteria, purple = fungi, and organ = synthetics. Above the timeline, discoveries are shown; below the timeline, resistance occurrence is presented. The picture was taken from M. Hutchings, A. Truman, and B. Wilkinson, „Antibiotics: past, present and future“, Curr. Opin. Microbiol [11].

The World Health Organization (WHO) has listed the most dangerous pathogens and has put categories on them based on their threat [3]. Categories are critical, high, and medium [2]. The list of critical bacterial species contains *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Klebsiella pneumonia*, and *Enterobacter* spp [9]. All four belong to the ESKAPE strains, which cause ¾ of all resistance deaths together with *Enterococcus faecium* and *Staphylococcus aureus* [6], [8]. Those two, *Enterococcus faecium* and *Staphylococcus aureus,* are united with *Helicobacter pylori*, *Campylobacter* spp, *Salmonella* spp, and *Neisseria gonorrhoeae* in the category of high threats. Medium threats are *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Shigella* spp [2]. Those pathogens have acquired so many different resistant genes that even last-resort antibiotics fail, and people die again from superficial infections. Alternative treatments are being tested and going into clinical trials, such as probiotics, antimicrobial peptides, antibodies, or phages. [6], [8]. But in this thesis, we will focus on phages only – the natural enemy of bacteria.

1.1 Targeted Single Phage Isolation

1.1.1 Phage History

Enormous efforts are made to eliminate bacteria during infections and to find suitable alternatives, as the number of newly developed antibiotics has been increasingly low over the last decades.

One suitable alternative is phage therapy, a therapy that has already been used to treat bacterial infections in the early days, even before antibiotics were developed (the history timeline of phage therapy is in Figure 2) [3], [7],[6], [12]. In 1915, phages were first found by Federick Twort but two years later again by Félix d'Hérelle [6], [13], [14], [15]. Félix d'Hérelle called them bacteriophages, which means "bacteria eater" [13]. They discovered that phages have the ability to destroy bacterial cells. Phages are viruses that need a bacterial cell to replicate and produce progenies. Phages are parasites and cannot survive without their hosts [6], [13]. When Félix d'Hérelle discovered that phages destroy bacteria, he investigated further and wrote his first paper about phages that decreased the bacterial load in dysentery patients [13]. Within rabbits, he evaluated the phage functions against *Shigella* and found that phages protected the rabbits. In 1930, companies started to produce phage products to treat bacterial infections broadly [13]. However, later that year, the Council on Pharmacy and Chemistry of the American Medical Association raised their concern as the outcomes of phage therapy are not as clear, and further research is necessary [13]. This was precisely the same time when antibiotics were discovered. Phage therapy experienced a pushback, and people lost interest. Especially Western countries lost complete track of phages as all their efforts were put into expanding their research on antibiotics. Only Eastern Europe and the Soviet Union kept using phages as a bacterial treatment, and the research [6], [14], [15]. The timeline of phage research is shown in Figure 2.

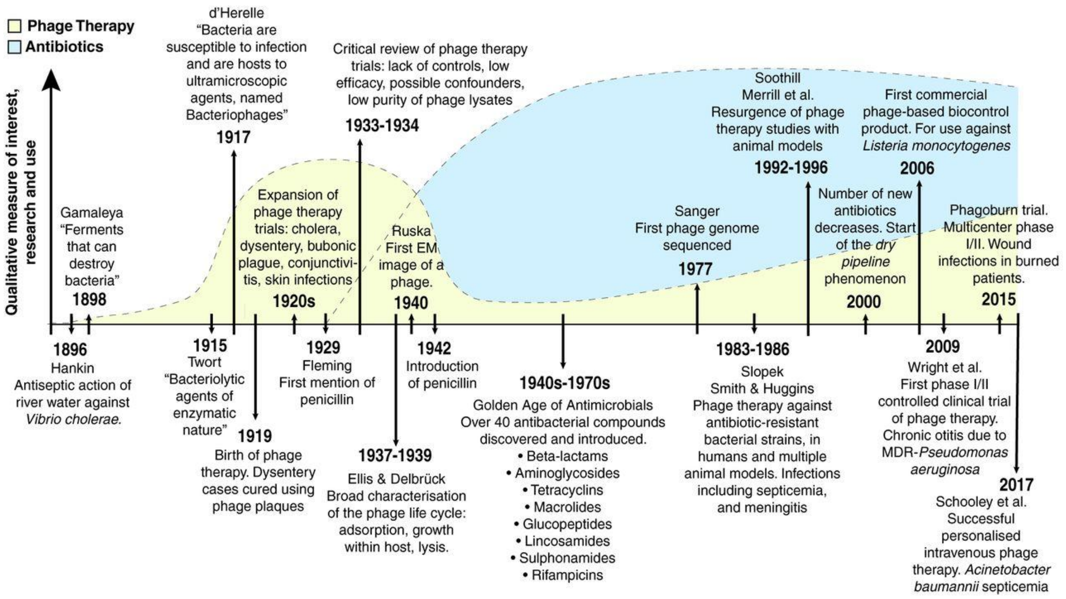


Figure 2_Timeline for Phage Therapy History

The evolution of phage therapy is represented by the yellow areas. Phage therapy kick-started around 1915. The blue area shows the development of antibiotics, which were accidentally found in 1929 by Fleming and changed infection treatment drastically around 1942. Picture taken from F. L. Gordillo Altamirano and J. J. Barr, „Phage therapy in the post-antibiotic era"[16]

1.1.2 Phage Biology

Fundamental phage research was scarce and limited since most experiments were conducted with *Escherichia coli* and its phages. However, the renewed interest expanded the field of phage research and constant new findings are shared. So, what are phages?

1.1.2.1 Phage Life Cycle

Bacteriophages are viruses that infect bacterial cells and hijack the host cell's machinery to reproduce themselves [3], [13], [15]. This process is their life cycle and is described in short: infection – hijacking – progeny production [17]. Phages are categorized into two major groups based on their life cycle: lytic and lysogenic [6], [18]. A third and less common life cycle is the chronic one. All three cycles are pictured in Figure 3. Phages that enter the lytic life cycle infect their host, reproduce themselves, and release progenies [14], [15], [17]. Those progenies can only be freed when two proteins are produced, holin and endolysin [19], [20]. Holin forms holes into the bacterial cell wall, making membrane proteins fully available for the endolysin, which then degrades the remaining wall. During that process, the bacterial cell is fully destroyed, and progenies are released [21].

The second life cycle is the lysogenic life cycle. The lysogenic life cycle is entered by those phages that infect the cell but then do not immediately kill it. Phages either integrate themselves into the host genome and become a prophage, or their genomic material is freed and floats as a plasmid [6], [14]. These phages can stay dormant for an extended period of time until a stressor or other influencing factors push them back into the lytic cycle, where phages replicate themselves and produce progenies [3], [15], [18], [22]. The pushback, however, is only possible as long as the genetic information is intact and complete; if parts are impaired or are missing, the induction can not happen anymore.

In both life cycles, lytic and lysogenic, phages destroy their host when progenies are released, which is the major difference compared to the chronic life cycle. Phages with a chronic life cycle infect their host, but while releasing progenies, they do not destroy their host cell [14], [17].

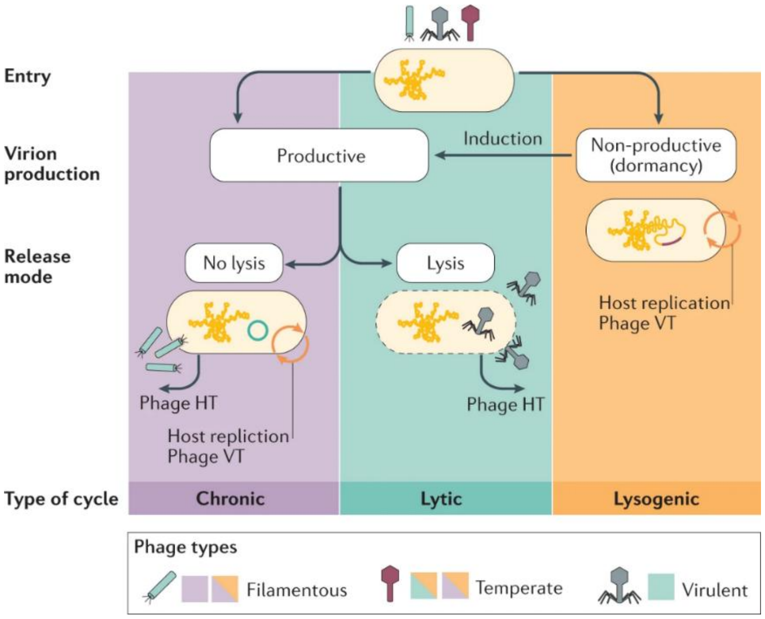


Figure 3_Phage Life Cycles

The picture shows the three life cycles for phages. The lytic and chronic life cycles reproduce progenies, whereas only the lytic cycle destroys the host during progeny release. Phages that enter the lysogenic life cycle do not produce any progenies at first. They integrate their genomes into the host genomes and stay dormant. However, temperate phages can be triggered to switch into the lytic cycle to generate progenies and destroy the host. Picture taken from A. Chevallereau, B. J. Pons, S. van Houte, and E. R. Westra, „Interactions between bacterial and phage communities in natural environments“ [23]

The majority of known phages go through either the lytic or lysogenic life cycle. Phages decide quickly after the adsorption which cycle they choose. So, the decision is made right during the infection process and made up before any other stage starts [15], [22].

The process is triggered by signaling molecules [18]. Signaling molecules are produced during the lytic cycle. As the concentration rises, the chances increase drastically for a cycle switch. If the concentration of the lysogenic-promoting proteins decreases, the likelihood of a cycle switch decreases with it. Recent studies have shown that the small molecule arbitrium is such a communicating molecule for *Bacillus* phages [15]. With its rising concentrations, phages are more likely to choose the lysogenic life cycle rather than the lytic one. The decision is based on which cycle is currently more beneficial for the phage. Environmental factors such as host availability or external stress factors play a role in their decision-making [22]. Supposedly, nutrients are rare, and the amount of phage relatives is too concentrated. In that case, phages tend to choose the lysogenic cycle as their chances for survival are higher as more progenies can be produced in a later stage. Additionally to environmental factors, if a host cell has multiple phages attached, all phages together decide in which life cycle it will be continued further [15]. The life cycle decision is complex, and only for phage lambda really well studied and better understood [15]. However, for the remaining phages, it remains a mystery.

1.1.2.2 Phage Properties

Phages can be categorized by their life cycle, but there are also other factors, such as their genomic information. Phages can have DNA or RNA, double-stranded (dsDNA/dsRNA) or single-stranded (ssDNA/ssRNA) [15], [24]. Phages can also be classified based on their morphology. Figure 4 presents the different morphotypes of phages. Phage research accelerates, and more and more details are added as fundamental research continues and new information is revealed. Currently, the majority of phages belong to dsDNA-tailed phages, which are also the most studied ones, but more studies are coming in, filling the research gaps around them [6], [15], [24].

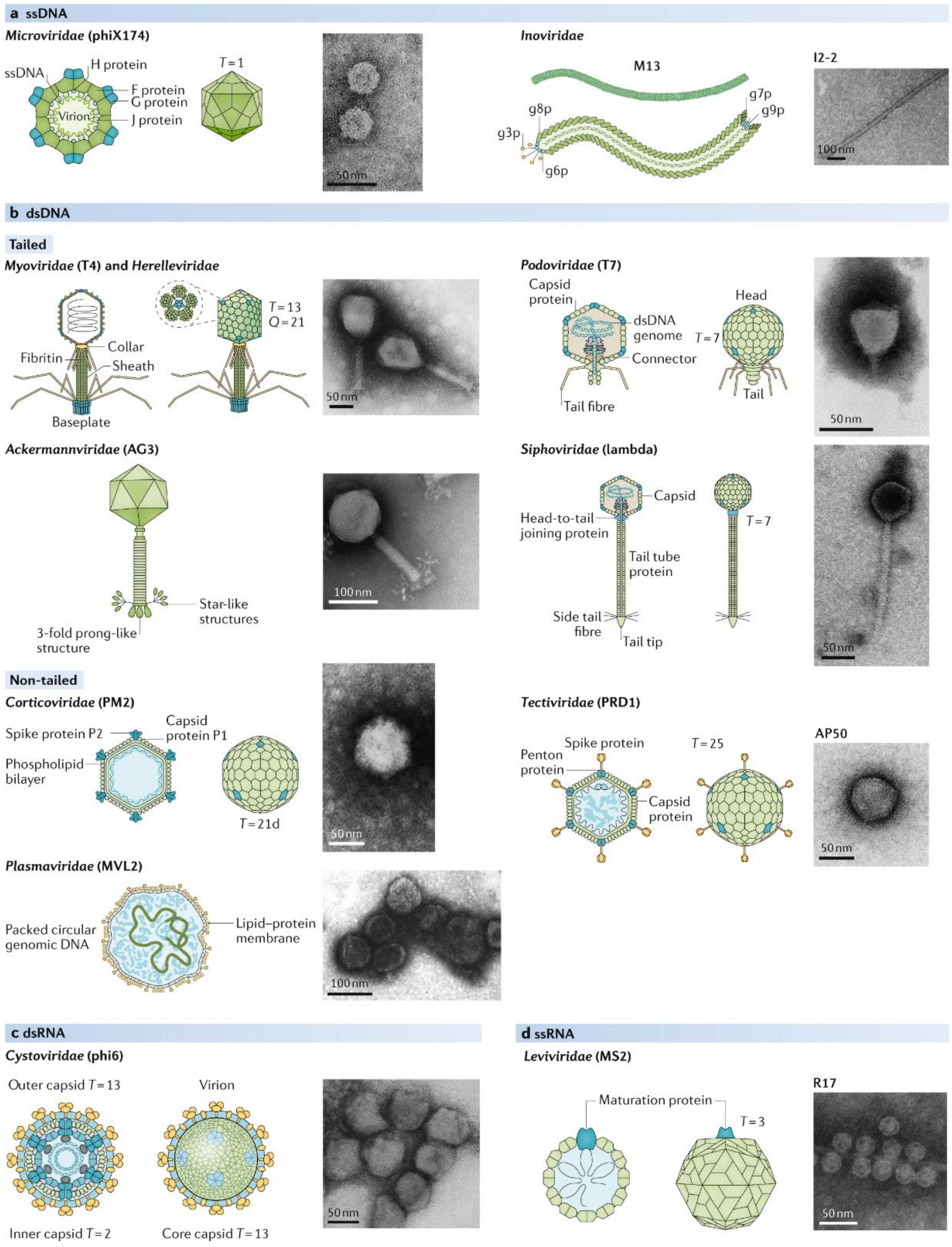


Figure 4_Different Phage Morphologies.

Picture was taken from M. B. Dion, F. Oechslin, and S. Moineau, „Phage diversity, genomics and phylogeny“, [24]

Researchers found that phages are amongst the most abundant species on earth, outnumbering bacterial cells by at least a factor of ten [14], [24]. 1 to 10 is the ratio in the gut, for example. The ratios can be smaller, especially in extreme environments such as the deep ocean. However, in general, scientists claim that phages are present wherever bacteria can be found [14]. Additionally, once phages are present, phages can persist for a long time as long as no disturbing factors occur. Disturbing factors are UV light or harsh chemicals that can influence or completely destroy the phage. If well protected, some phages can survive for a long time, even for decades – often stabilized by a high concentration [14]. However, others are more fragile and sensitive and break down very easily, even without external help. In general, phage survival relies heavily on the phage itself and cannot be generalized.

Despite their abundance, phages can also be differentiated based on their biogeography [14]. That means that not every phage is located everywhere in the world; some phages are only found in closed surroundings, whereas others, such as the crAssphage, can be found worldwide. The crAssphage is part of the gut microbiome and is present globally [15], [25].

Another differentiation factor is their genome. Bacteriophages are very different compared to their bacterial hosts. The phage genome size ranges from 2 kbp up to 500 kbp; the latter is then called Jumbo phages [24]. The genome does not include any universal marker genes for genotyping, like the 16S rRNA gene in bacterial cells, and their genetic information is highly dispersed, more like a mosaic [26]. Their genome complexity makes phages highly diverse, but phages increase their diversity even more than similar genetics do not equal similar behavior. Phages can have a big portion of their genomes identical, but their behaviors differ completely. Behavioral differences have been seen for phages, which have nothing in common [24].

Another layer of diversity is added by their morphology, as pictured above. Phages can have packed their genetic information in a polyhedral capsid, a common geometrical form in viruses. Their genetic information can also be stored in a filamentous capsid, which is tubular. Besides their capsids, phages can also have tails. The tail can then either be contractile or not. Phages that have a polyhedral capsid and a contractile tail belong to the family of *Myoviridae*. If the tail is non-contractile, phages are identified to the *Siphoviridae* family, and no tail classifies them as *Podoviridae* [27].

1.1.2.3 Phage Taxonomy

In the early days, when phage research was limited, their classification was based on their morphology and genome types, such as dsDNA, ssRNA, or host range. In 1978, the International Committee on Taxonomy of Viruses (ICTV) approved the phage families *Inoviridae*, *Microviridae*, *Tectiviridae*, *Corticoviridae*, *Plasmaviridae*, *Leviviridae* and *Cystoviridae* [28]. Twenty years later, in 1998, the order Caudovirales was approved, which combined all tailed phages and included *Myoviridae*, *Siphoviridae*, and *Podoviridae*. However, since phage research rapidly increased and sequencing technology advanced, more and more insights about phage genomics were gained. Consequently, the new information quickly made the used taxonomy irrelevant as there was no cohesiveness anymore. Therefore, phage taxonomy is now undergoing major work [28]. Especially the order *Caudovirales* is highly affected as the majority of sequenced phages are tailed phages. So, *Myovirus* now includes three families: *Ackermannviridae*, *Chaseviridae*, and *Herelleviridae*. The *Siphovirus* family combines *Demerecviridae* and *Drexlerviridae*, and the *Podovirus* family consists of *Autographiviridae*. This order was only the beginning, as the ICTV is constantly working on expanding and categorizing existing and newly found phages [28].

1.1.3 Phage Innovations

Not only did taxonomy need to take in a lot of new information, but the entire phage world was getting swamped with novel research outcomes. In the early days, phage research was limited to a small number of known phages such as T1, T4, lambda, T7, and phi29, which are mainly *Escherichia coli* phages, except the latest. Phage knowledge was limited, but now, phages are used in a wide range of biotechnological tools. One tool is phage display. In phage display, filamentous phages are used to express different surface proteins to catch the interaction between antibodies or other proteins [29]. A plethora of biotechnology tools was available after the discovery of molecular cloning. The discovery of molecular cloning and gene expression systems was embossed by bacteriophage lambda. Lambda was the center of attention in genetic engineering between the 1950s to 1980s [30]. Genetic engineering even won the Nobel Prize. The Nobel Prize for Chemistry in 2020 was awarded for genome editing using the Clustered Regularly Interspaced Palindromic Repeats (CRISPR) Cas9 system. CRISPR Cas9 is a bacterial defense system against bacteriophages and can be used to delete/add/change entire genes within the genome [31]. Another gene editing tool is the Cre-Lox system. The Cre-Lox system is nowadays used for site-specific recombination enabling scientists to modify genes [15]. A major revolution was the discovery of the phi29 DNA polymerase. Phi29 DNA polymerase changed the world of biotechnology as single-cell amplification was possible, and PAcBIO sequencing was developed [15]. Phages are also used to detect bacteria in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOFMS). Bacterial concentrations are often below the detection limit, but scientists discovered that they can identify bacteria indirectly. They mix the bacteria with phages, and during the amplification process, certain proteins are produced that are specific to their targets. Those proteins can then be detected in the MALDI-TOFMS and confirm the presence of their target bacterium [32]. The technique detects the bacterium indirectly based on phage biomarkers produced during the amplification process. But the greatest phage feature is that they can destroy bacterial cells and support our battle against antibiotic-resistant bacteria.

1.1.4 Phage Detection and Isolation

Phages have to be found first in order to be used as an anti-bacterial agent. They are commonly isolated from sources where the infecting bacteria are suspected to live in [8]. Recently used phages were found mainly in environmental samples such as hospital sewage, stool samples, patient samples, or soil. Phage isolation/detection is a challenge despite their abundance. There is currently no standard operating procedure available for phage detection or isolation, and to generate a solid picture for one phage, an entire array of tests needs to be performed. Phage quantification can be done via Epifluorescent microscopy or flow cytometry [14]. With a transmission electron microscope, the phage`s morphology can be determined, and with metagenomic sequencing, the genome [14]. However, none of the previously mentioned methods can actually tell if the phage actively infects the bacterium or not [14]. These techniques are purely detection only. Active infection can only be determined via phage isolation. Phage isolation can be done either by spot tests or a double-layer plaque assay. Spot assays have a lawn of bacteria on an agar plate, and viral sources are spotted right onto them, incubated, and screened for lysed areas. Double-layer plaque assay starts with an amplification process first. The target bacterium is incubated with the viral source. On the next day, the supernatant is separated from the bacteria, and freshly grown bacteria are again incubated with the supernatant and fresh media. These culturing rounds are repeated multiple times. Between the rounds, the supernatant is tested for phages via a double-layer plaque assay. The target bacterium is mixed with supernatant and soft agar (0.7 % agar), poured onto an agar plate, and incubated. After incubation, clear lysis zones confirm the presence of phages for that specific bacterial strain [33]. To ensure that only one phage is isolated, purification rounds need to be performed. For that, plaques are picked and incubated with the target bacteria. After incubation, the supernatant is tested again with a double-layer plaque assay. Usually, three rounds of purification are executed. Purification is extended if plaque morphology keeps changing or multiple different plaque sizes are still visible [33]. After purification, phage characterization can start.

1.1.5 Phage Characterization

Phage characterization is a fundamental element in phage research. Only fully characterized phages can be administered to patients in phage therapy [8]. The essentials are origin, family, subfamily, morphology, life cycle, host range, potential toxins, virulence factors, antibiotic resistance, stability of lysis, frequency of resistance bacteria to phages, and temperature optimum. All these factors must be investigated, which is a lengthy and laborious task [12], [34]. Additionally, to the before mentioned essentials, the latency period and the burst size must be determined as well. These properties are important factors in phage therapy [33]. In phage therapy, phages should be quick within their replication cycle and produce progenies effectively, which means having a short latency period (time of attachment until lysis) and a high burst number (numbers of progenies) [8], [12], [34]. Phages should also follow the lytic life cycle as prophages could transfer potentially virulent or toxic genes to the host and jeopardize the success of phage therapy [3], [12].

1.1.6 Phage Therapy

Phage therapy is a method to fight bacteria that can be accomplished with two strategies: a targeted approach or a broad one [8]. The broad approach is similar to a broad spectrum-antibiotics [33]. The broad approach contains a phage cocktail that targets a wide range of bacterial species. The cocktail holds an array of different phages and is available quickly since cocktails are often premade [8]. In comparison to the broad approach, there is the targeted phage approach. The targeted cocktail is firstly not premade and, as the name says, targeted. The targeted cocktail is directly shaped for the infecting bacteria as phages are only isolated for that patient with this specific bacterial infection. However, the targeted approach is only available on request, whereas the cocktail with the broad approach can be purchased in pharmacies in Poland, Russia, and Georgia [8]. These premade cocktails are made for specific infections but not for specific strains. Premade cocktails are not available in Western countries, as phage therapy is not legal and permitted yet. Currently, if patients are treated with phages, it is under the umbrella of compassionate use (Declaration of Helsinki) or expanded access, with the patient`s consent, and considered experimental treatment [8]. Administered phages are magistral phages that are produced by pharmacies (in the scope of normal pharmacy operation) or by physicians (direct use on patients) [10]. Magistral phage products can be individualized and are more patient-directed than over-the-counter phage cocktails. Phages are currently unauthorized medical products that will only be given in life-threatening situations with no alternative treatment options [10]. Phage therapy is currently not accepted by the Food and Drug Administration (FDA) or the European Medicines Agency (EMA) [35]. Additionally, each country handles phage therapy with compassionate use differently, as laws and medical regulations vary.

Phage Therapy Advantages

Phage therapy has many advantages. Phages are self-dosing. As long as bacterial cells are present, phages can replicate [12]. If the targeted pathogen is eliminated, phages will not attack randomly others and, therefore, cannot reproduce themselves anymore. Overdosing is consequently not possible. Additionally, an overdose of phages does also not harm the patients as long as all toxins (e.g. lipopolysaccharide, LPS) are removed during the production process. Those toxins spike the immune system and can result in a septic shock, the same as in bacterial-caused sepsis [12].

A second advantage is their biofilm-degrading ability. If the right phage is found, phage therapy can still be carried out even in the presence of biofilms [3]. Phages can break down biofilms by producing enzymes that can degrade biofilm polymers. Over 160 depolymerases have been identified, separated into two classes: hydrolases and lyases. Those enzymes disrupt biofilm structures, deconstruct them, and free the path to the inner core, exposing bacterial cells.

Exposed cells are then available for phage infection and lysis. This is a major advantage in phage therapy, that phage administration is still possible despite the biofilm.

Another advantage is phage specificity. Phages are strictly specific to their host and, in the majority of cases, will not infect randomly any other species [6], [12]. Some phages infect only one strain, whereas others can infect multiple different ones. Phage isolation and infection only work with the right hosts. For example, it is highly unlikely to isolate a phage for a gram-negative strain using a gram-positive strain [33]. Additionally, phages have little to no effect on mammalian cells. Their natural hosts are bacterial cells [3].

Phage Therapy Disadvantages

Phage therapy also has its disadvantages. One major disadvantage is the lack of standardized protocols. There are currently no standards on how to isolate/process phages, which consequently results in a no-GMP production for phages [8]. However, despite all the odds, current testing procedures are reliable and accurate, with one major drawback. They are not fast [35]. Determining the best phage means going through phage banks or finding the needle in the haystack by using environmental sources and conservative techniques such as spot tests or double-layer plaque assay [12]. Spot assays or double-layer plaque assays are easy but time-intensive (as their incubation is mostly overnight) and laborious. Another challenge in phage isolation is that phages can behave differently from host to host. Their burst size can change, their latency period might differ, and so can their temperature optimum for infection. Another problem is that some phages might not be able to form any plaques as the agar is too dense for plaque formation. Bacteria might also not grow in agar, making phage isolation and characterization difficult or, in some cases, impossible [33]. One major characterization step is determining the phage's host range. The host range must be tested on each strain separately. Acceleration can be achieved by testing multiple phages via spot assay on one plate. However, a double-layer plaque assay must be executed to ensure that the phage is actively infecting this host. Multiple different concentrations can also help as sometimes bacterial cells are killed by unknowns in the viral source but not actually from the phage itself [34].

Another disadvantage is the rather biased process of phage isolation. The traditional isolation technique includes an enrichment step [33]. That step might favor phages with a higher turnover rate and exclude phages that are slow (low latency period) or with a low burst size. The technique also includes a filter step to eliminate all bacterial remains, but some phages are big and can be lost during the process. Those losses limit the choices for phage therapy. A variety of phages is crucial for phage therapy, which is currently not accomplished due to the techniques used [33]. Also alternative techniques such as liquid assays are insufficient. In liquid assays, the bacterial culture is incubated together with the viral source, and optical density is measured. If the culture has cleared, phages with lytic ability are highly likely [33]. A major drawback of that technique is that there is no differentiation between one single phage and infection with multiple different phages. The differentiation producer is only possible via double-layer plaque assay. A double-layer plaque assay can distinguish between different phages, as plaques might differ in their appearance.

More advanced tools like bioinformatic analysis can support and clarify, but only to a certain extent. Computational science can predict the interaction between host and phage, which might help to find the right phage from the phage databank [35]. However, as most phage sequences are incomplete or fragmented, the predictions can be inconclusive, and in-vitro tests have to be executed [35]. Already existing data can also support the hunt to find the right phage for phage therapy. Standards are consulted from well-studied model phages to predict possible outcomes from newly isolated phages. Such model phages are Dp-1, T4, T7, MS2 or phiX174 [8].

All these ideas and modifications helped phage therapy over the last few years to avoid the long and exhausting process of finding the right phage. But, if the first hurdles are tackled, such as isolation and characterization, others quickly pop up, such as pharmacokinetics. Pharmacokinetics describes the effect of the medical agent within the patient's body. The agent would be phages in the case of phage therapy. However, phages are dynamic and describing the effect of phages in the body and their actions are complex. Phages are big, their diffusion rate is low, and the immune system will likely attack them since phages are foreign to the patient`s body. Consequently, the actual phage concentration is lowered by a hundredfold when reaching the target site [12]. This reduction is even worse when phage cocktails are applied, as the starting dose in cocktails is already lower for each phage than the concentration of a single phage. Another disadvantage of phage cocktails is the interaction of phages within the cocktail which is complex. As a consequence, phage therapy is often better applied directly at the infection site to get the right concentration [12].

As these hurdles are not enough, phage therapy has a major problem, as pathogens can become resistant to phages. The rapid evolution in better sequencing technology has boosted phage therapy and phage research in general, which revealed that bacteria can become phage-resistant quickly and easily [12]. Often, a tiny bacterial surface change can suffice that the phage cannot attach to its host anymore and become useless. If phage resistance occurs, it does occur in the entire bacteria population and not only in one phage-host pair [12]. The likelihood of phage resistance is prominent during treatment. Phage resistance occurs in many ways. Bacterial cells have an entire set of defense mechanisms that they can use to get rid of the phage or avoid phage infection. The most heard one is CRISPR Cas9, but bacteria also have restriction endonucleases that cleave phage DNA or abortive systems to kill themselves before the phage can replicate. However, since phages and bacteria co-evolve constantly and have been over the years, phages have also developed existing strategies for bacterial defense systems [6], [12]. This co-evolution can happen quickly and even within one infection cycle. The rapid evolution of phages is a big advantage compared to antibiotics, which are chemically and structurally rigid and need scientists to overcome resistance by creating a totally new antibiotic. Ideas to overcome resistance are available, as phage resistance is an issue in phage therapy. There are certain ways to fight resistance, except phage evolution. Patients could get a phage with a broader host range [12]. A broad host range might help, as phage might use different receptors for the infection, and the pathogen has only energy to defend itself on one infection route. A second way would be fast phages. Phages that can kill off their host quickly and release a huge amount of progenies can reduce the risk of resistance as there is no time for defense adaptation [12].

A third approach is phage cocktails to avoid phage resistance. Phage cocktails contain multiple different phages with different infection routes, but all of them have the same target. Due to the attack of many different phages, the host is less likely to develop resistance [12].

With all the acquired knowledge, scientists went into clinical trials with phage therapy.

Clinical Trials

Phage therapy was not a success story in clinical trials, despite their promising traits and theoretical evidence and proven individual patient stories where antibiotics failed [3], [7]. https://clinicaltrials.gov lists currently 43 clinical trials for the search phrase "phage therapy". 41 of those trials are phage therapy trials. The details are listed in Supplementary Table 3. Clinical trials are necessary to advance phage therapy for clinical development. A major task will be to update and adapt regulations as current laws are made for synthetic drugs, which do not change their structure while being administered or being on-site for infection. Whereas phages do self-regulate themselves, replicate at the infection sites, and are capable of changing their genetics while being inside the patients [3]. Phages are less controllable inside the body than drugs, which classifies them as biological agents and need different regulatory settings compared to conventional drugs [3].

Phage therapy is further experienced in livestock and mammalian farms than in humans, as treatment regulations in livestock and on farms are less rigid than for human patients. Bovine diarrhea, bovine mastitis, pig diarrhea, and many other poultry diseases have successfully been treated with phages, and the first over-the-counter phage medications are allowed in the US and Ukraine [3]. Over-the-counter phage cocktails (also for humans) are also available in one of the oldest phage therapy centers, the Eliava Institute in Georgia. The Eliava Institute provides phage therapy since 1923. Since 1970, phage therapy has also been available at the Hirszfeld Institute of Immunology and Experimental Therapy in Wroclaw, Poland, and since 2005, within its own phage unit. However, in the Polish phage center, phage therapy is only available under the umbrella of compassionate use, as phage therapy is not accepted yet in the European Union.

Every phage therapy case in that center complies with the Declaration of Helsinki and the Medical Profession Act of 5th December 1996 of Poland [36]. In January 2018, Belgium changed its law and allowed the treatment of bacterial infections with "magistral phage medicine". These phage products are made by pharmacists for physicians only when patients need them. Later that year, in June, a phage center was opened in San Diego, providing phage therapy with the allowance of the FDA [36]. More and more phage therapy centers pop up often in collaboration with universities such as the University of Leicester, UK, or Yale University, USA. Additionally, to all phage centers, a community is forming around the globe, making phages accessible to everyone who needs them, and biotech companies are founded, such as biomX or PhagePro.

1.1.6.1 Phage-Derived Antibacterials

Despite all the newly ignited interest in phage therapy, scientists also expanded their research field to phage-derived antibacterials. Phages produce enzymes, which help them to break down bacterial cells [3]. Those enzymes are chemical molecules such as antibiotics, which can be easier to put through access regulations. Endolysins, which are such enzymes, are highly sufficient in lab conditions but currently fail in clinical trials. Under lab conditions, endolysins work perfectly on gram-positive bacteria yet not so much on gram-negative pathogens due to their outer membrane [3]. Another disadvantage of phage derivates is that experiments have shown that they only work shortly after infection or even only as prevention. Additionally, phage enzymes cannot be delivered to inner organs as they are detected as foreign and being attacked by the patient's immune system, which can lead to full eradication of the enzyme before hitting the target site [3].

1.1.6.2 Phage – Antibiotic Combination

Another aspect of phage therapy is phage-antibiotic combination therapy. Phage-antibiotics combination therapies have shown tremendous success. The success is based on the changes in the defense system of the pathogen [37], [38]. Antibiotic-resistant pathogens spend their energy on becoming resistant to antibiotics, but in the event of a phage infection, pathogens need the energy to escape the phage. Therefore, the energy is used to adapt the defense system against the phage and in that process, they often lose resistance against the antibiotic. It is a trade [37], [38]. These trades can happen in various places. Surface receptors can be altered to avoid phage infection, which then changes the LPS. The new LPS structure could then be more sensitive to antibiotics [37], [38]. The alterations can also happen in the cell membrane proteins, such as drug-efflux transporters. Phages can use those transporters as attachment points. If the pathogen wants to get rid of the phage, it needs to fully change those transporters via DNA changes or eliminate them completely. The elimination of the drug-efflux transporter wiped out the antibiotic resistance and with that, the pathogen is sensitive to antibiotics again. In these cases, either the phage or the antibiotic will destroy the pathogen [37], [38]. In other cases, phage-antibiotic combinations are favored as phages are used to eliminate the biofilm prior to antibiotic usage. Many pathogens hide behind a biofilm and become unreachable for antibiotics, but phages have biofilm-degrading enzymes that can destroy the biofilm and can ensure that the antibiotic reaches the cells [37]. Current data are experimental. However, first results show a synergistic effect when the antibiotic is chosen with care [3]. For combination therapy, it is highly important to understand the antibiotic`s mode of action as it could interfere with the phage and lead to the total opposite result. In general, phage-antibiotic combination therapy also needs further investigation and broader experimental setups to fully comprehend the situation [3], [39].

1.1.7 Objectives

As phage therapy progresses, more characterized phages are required to meet the demand and diversity of all patients. As traditional techniques fall short of that, novel techniques need to be developed. In the first part of my thesis, I aim to create a method that isolates phages fast and covers the entire phage diversity for the targeted strain. Additionally, I want to incorporate a characterizing step to identify potential candidates quickly. This should help to save time when time is of the essence. Additionally, the method should also be adaptable for all common pathogens, gram-negatives, and gram-positives and work anaerobically.

1.2 *Helicobacter pylori* Phages and Their Diversity

1.2.1 Helicobacter pylori – the Bacterium

Phage isolation becomes an even greater challenge when the chosen pathogen is a challenge itself. *Helicobacter pylori* (HP) is such a challenge. *Helicobacter pylori* is one of the most spread pathogens, with an infection rate of 50 % of the world's population [40], [41], [42]. It has spiral-shaped cells that are microaerophilic and composed of a negative-gram bacteria cell wall [41], [42]. Its natural habitat is the stomach, as it adapted to the harsh, acidic environment. The stomach acid is neutralized with the urease enzyme, and its flagella help to move through the mucus layer into a less acidic surrounding. If *Helicobacter pylori* remains in the stomach, it can change its appearance from spiral to coccoid shaped – a form that enables its survival. *Helicobacter pylori* transitions to the coccoid state whenever stressors are present [43], [44]. In coccoid form, *Helicobacter pylori* can stay dormant and survive drastic environmental settings, but as soon as it arrives back in its normal habitat, it changes back to its living form. In this coccoid form, *Helicobacter pylori* is also immune to antibiotics and cannot be cultured in laboratory settings anymore [45]. Another unique property of *Helicobacter pylori* is its genomic composition. Its genome is highly unstable and varies a lot [46]. The strains differ from each other by rearrangements, inversions, and deletions. The genomic sequence of a strain can also change during an infection. The reasons for that are transposable elements, restriction and modification enzymes, and a bad DNA repair system [46]. This kind of genetic flexibility has been studied well in the two model strains, *H. pylori* PMSS1 and *H. pylori* SS1.

1.2.2 *Helicobacter pylori* PMSS1 and SS1

Helicobacter pylori PMSS1 (pre-mouse Sidney Strain) and SS1 (Sidney Strain) are two highly prominent model strains. *Helicobacter pylori* PMSS1 was isolated from a patient in Syndey, Australia, from a gastric tissue biopsy. SS1 was created from PMSS1 by scientists via subculturing the strain in human gastric homogenates and infiltrating the strain into a mouse gut [46]. Eventually, SS1 adapted to the rodent environment and is now infecting mice. As SS1 derives from PMSS1, both strains are highly similar and have key virulence factors such as VacA, CagA, cagPAI, and TSS4. Yet, SS1 does not have a fully working cagPAI TSS4 as cagY is defective [46], [47]. Whereas PMSS1 has this pathogenicity island fully functioning and can consequently infect humans and mice. Additionally, SS1 lacks two genes that are present in PMSS1: starvation protein A (CstA) and Lactate permease (LldP) [46]. Since SS1 was created from PMSS1, the *Helicobacter pylori* pangenome was not extended. Those two strains have a 99.9 % identity. One difference occurs on the plasmid, and 46 differences are between them on the genome level - 28 in RNA or protein-coding regions and 18 in intergenic regions [46]. Another difference is that a ¼ of the SS1 genome are inversions of PMSS1 and run opposite directions. However, those diversions can also run in the same direction as the genes in PMSS1. These genetic traits are unique and vital for *Helicobacter pylori* and are key elements during infection.

1.2.2.1 Infection Routes and Rates

Helicobacter pylori infections are quite frequent. Its transmission routes are either oral-oral or fecal-oral, and infections occur highly likely within family settings or shared living spaces. Poor hygienic standards increase the risk of infection, and so do poor socio-economic situations [48]. Correlations are seen in less developed countries, where infection rates are up to 70 % in some areas, whereas Switzerland has only 19 %. *Helicobacter pylori* is detected either via Urea breath test, antigen tests from stool, gastric biopsy, or cultivation [48]. Each method has its own advantages and disadvantages with regard to sensitivity or costs.

1.2.2.2 Treatment Plan

If the tests are positive, WHO suggests a full eradication therapy. Despite the fact that 90 % of the positively diagnosed people have never experienced any symptoms as their infection remains asymptomatic [44], [48]. However, for the remaining 10 %, a *Helicobacter pylori* infection can cause duodenal or gastric ulcers (1-10 %), gastric carcinoma (0.1-3 %), or mucosa-associated lymphoid tissue lymphoma (MALT - <0.01 %) and 780,000 infected patients die each year globally [42], [48]. Consequently, *Helicobacter pylori* should be eliminated completely, preferably with antibiotics. Antibiotics have drastically minimized *Helicobacter pylori* infections since the pathogen was discovered 40 years ago [49]. The standard therapy usually consists of three drugs: two antibiotics and a proton pump inhibitor. Therapy with an additional substance, bismuth, has also been tried [42]. However, due to its nature, *Helicobacter pylori* quickly became resistant to antibiotics – especially to clarithromycin, levofloxacin, and metronidazole [50]. If antibiotics are administered now for treatment, physicians choose them now based on the resistance rate within the administered region. Antibiotic resistance rates within *H. pylori* are skyrocketing and are highly dependent on the area of occurrence [48]. And despite its massive resistance, *Helicobacter pylori* is still treated with antibiotics since alternative treatments are currently unavailable. Alternatives such as vaccines are not ready yet or others struggle as treatment options since other obstacles such as low pH, high bacterial load, and impaired mucosal wall must also be considered [48].

1.2.3 *Helicobacter pylori* Phages

One alternative is phage therapy. However, not much is known about *H. pylori* phages. Around the same time when *Helicobacter pylori* itself was first found, its phages were detected by Marshal and Goodwin as well [45], [51], [52]. In 1990, the first findings were published about phage particles that were spontaneously generated. Only three years later, scientists discovered the lytic cycle of this phage, and the electron microscopy revealed that its head was about 50-60 nm [45]. Its genome size was estimated to be 22 kbp. With the revolution of sequencing techniques, more and more prophages were detected within *H. pylori* sequences – in almost 20 % of all isolates, they found prophages [45]. The first isolated prophage was PhiHp33 from the strain *H.p* B45 of a MALT patient. The prophage was induced via UV light [44]. Other prophages followed via UV light induction, such as HPy1R. HPy1R has a *Podovirus* morphology, double-stranded DNA, and a genome size of 31 kbp. HP1 phage was recovered from the strain SchReck290 with 22 kbp and showed similarity to phages from the *Siphovirus* morphology [51]. In Japan, two phages were detected via spontaneous induction: KHP30 and KHP40. KHP30 has a genome size of 26,215 bp, and KHP40 is 26,449 bp long. Both have a guanine-cytosine content (GC-content) of 35.8 % and are highly similar, with 96 % query coverage. Proteins such as integrases and primases were found, but the majority of the proteins were not identifiable [45], [53]. Another phage, 1961P, was found by screening 46 patients' samples and double-layer plaque assay. The newly isolated phage is 27 kbp long, has 33 ORFs (Open reading frames), and has a *Podoviridae* shape [54].

Despite all attempts and current known phages, the reference database of *Helicobacter pylori* phages is small and very limited in detailed information. Detection and isolation of *Helicobacter pylori* temperate and virulent phages were reported, but full characterization studies are missing [45].

1.2.4 Objectives

The number of failed *Helicobacter pylori* treatments increases due to antibiotic resistance, and alternatives are urgently needed. Since phage therapy is a successful alternative, scientists also investigate this option for Helicobacter pylori infections. Therefore, this chapter of my thesis is dedicated to finding new *Helicobacter pylori* phages, expanding the reference databanks, and deepening the knowledge about phages that we found via viral tagging.

1.3 Colorectal Cancer Cross-Infection Study

Another research field where little is known about phages and their role is the gut microbiota and its diseases.

1.3.1 Gastrointestinal Tract and its Microbiota

The gastrointestinal tract (GI) is one of the body's biggest surfaces and constantly interacts with the host, the environment, and the immune system. It is around 250 to 400 m^2 and exposed to roughly 60 tons of food in an average lifetime [55]. The existing microbiota, which consists of bacteria, viruses, fungi, archaea, yeast, flagellates, ciliates, and also protozoa, are then mixed with microorganisms from outside [26], [56], [57]. The gut has more than 10^{14} microbial cells, which is 10x more than human cells and is, therefore, the place with the highest number of cells [58]. Its microbiome (entire genome from all microbiota and their products) is 100x more than the human genome [55], [59]. The microbiota is a huge part of the body and, consequently, greatly influences its host. All microorganisms live in a mutually beneficial agreement, and by doing so, they support gut integrity, shape the intestinal epithelium, take part in the digestion process and metabolisms, and influence the immune system and overall health of the host [55]. If the agreement is disrupted due to an imbalance of bacterial concentration or diversity, it can lead to dysbiosis, which has a direct effect on the host's health [55]. Despite decades of research, the exact consequences of a balanced and dysbiotic state have not yet been revealed.

1.3.2 History of Microbial Research in the Gut

For many years, the gut microbiota has been full of mysteries. First insights were given into the gut microbiota in the late 17th century by Antonie von Leeuwenhoek when he did some experiments with the microscope [60]. Two centuries later, first cultivation studies were performed, enabling us to see microorganisms without a microscope. Despite all efforts over the years, many bacterial species, however, are still uncultivatable. An uncultivatable identification was possible when the first sequencing techniques, such as 16S rRNA sequencing and the first phylogenetic classifications, were developed. Novel sequencing techniques were developed in the late 20th century and revolutionized science [60]. Metagenomic sequencing brought more light into the dark matter of the gut with higher resolutions and higher sensitivities of the techniques [55]. Strain-level distinction was now possible. Additionally, multi-omics-technology enabled us to evaluate the transcription of proteins and look into the functional and metabolic states of the microbial ecosystem at specific time points and conditions [60]. Shotgun metagenomic sequencing revealed 1952 unclassified species where only 553 bacteria could be cultivated from the gut [59]. The gut's main components were *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Fusobacteria*, *Verrucomicrobia*, and *Bacteroides*. Whereas 90 % of the gut contains *Firmicutes* and *Bacteroidetes* [26], [56]. The historical timeline is visualized in Figure 5.

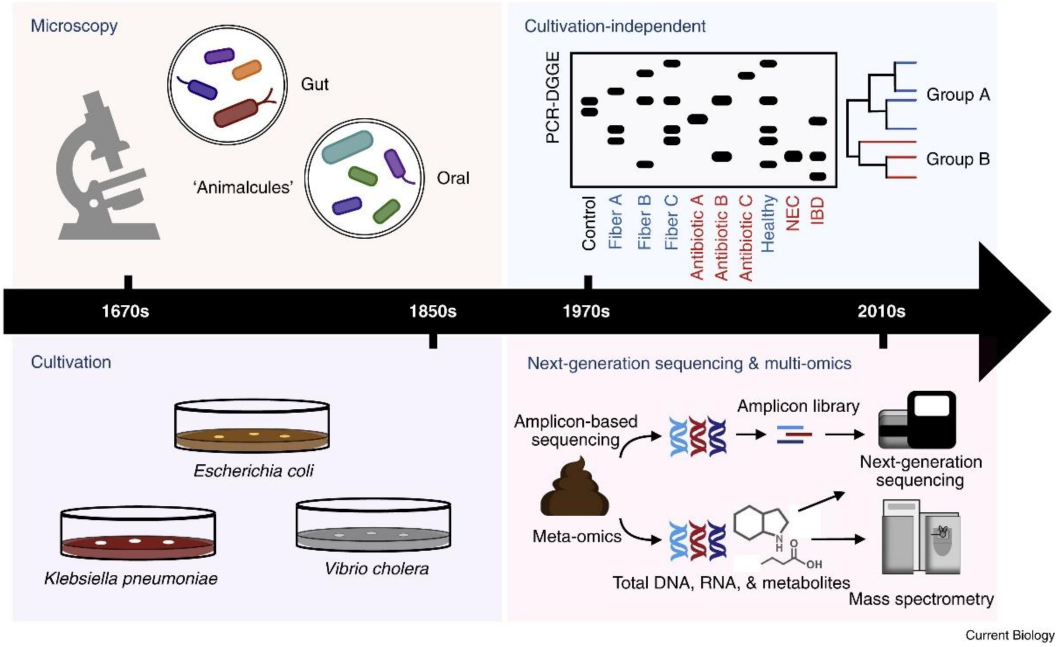


Figure 5_Timeline of Gut Microbiota Research

Gut microbial research was first conducted via microscopy and cultivation. Later on, new technologies were invented, such as PCR. With the development of new next-generation sequencing and multi-omics technology deeper insights were possible. The picture is from G. A. Kuziel and S. Rakoff-Nahoum, „The gut microbiome“[60]

1.3.3 Gut Microbial Colonization

After the initial identification of the gut microbiota, scientists asked who started the colonization. Scientists believe that the first gut colonization starts right after birth, and the first colonizers depend on the mode of delivery [55]. Vaginal-delivered babies have an increased load of *Lactobacilli*. C-section babies are slow with the colonization of *Bacteroides* but have a higher load of *Clostridium* species. It was also seen that vaginal-delivered babies have a more similar gut microbiota to their mothers than C-section babies. Within the first few months, the composition of the gut microbiota varies a lot but stabilizes itself around 2.5 years when it becomes similar to an adult microbiota [55]. Scientists have noticed that the gut microbiota is stable over a long period of time on a taxonomic level. On a species level, changes are more dynamic and vary [26]. However, the gut microbiota can be easily disrupted by antibiotics or other external factors. External factors are smoking, surgeries, depression, geographical locations, or general living standards. Researchers discovered that gut microbiota mirrors its outer surroundings – its living environment, its diet, and its host and is only slightly biased by the host's genes. Especially for infants, the gut microbiota can easily be shaped by food, and every change has a great influence on the infant's development [55], [56].

1.3.4 Gut Microbial Organization

It was also seen that not only the way of delivering and living surroundings influence the gut microbial composition, but also the area of the gut needs to be considered.

The gut microbiota composition and concentration differ depending on the area of the gut (see picture Figure 6) [26]. Which bacteria colonize the gut are dependent on many factors such as acidity, bile acid, digestive enzymes, antimicrobial proteins, chemical parameters, and the amount of oxygen, but physical parameters also play a role, such as peristaltic or gut structure [58], [60]. Starting at the top, the stomach is a heavily acidic area and is the least colonized area; only bacteria such as *Helicobacter pylori* can survive there. The small intestine has less acid than the stomach but is still highly acidic and oxygen-heavy. The small intestine is colonized with rapidly growing bacteria and facultative anaerobes, which can stick to the mucus / epithelial wall, such as *Lactobacillaceae*, *Firmicutes*, and *Proteobacteria* [55], [58]. The colon, in comparison, has a more diverse composition, more anaerobes, and bacteria that can break down complex carbohydrates, such as *Prevotellaceae*, *Firmicutes*, *Ruminococcaceae*, *Bacteroidetes*, *Actinobacteria*, *Proteobacteria*, *Akkermansia*, *Lachnospiraceae*, or *Rikenellaceae* [55], [58].

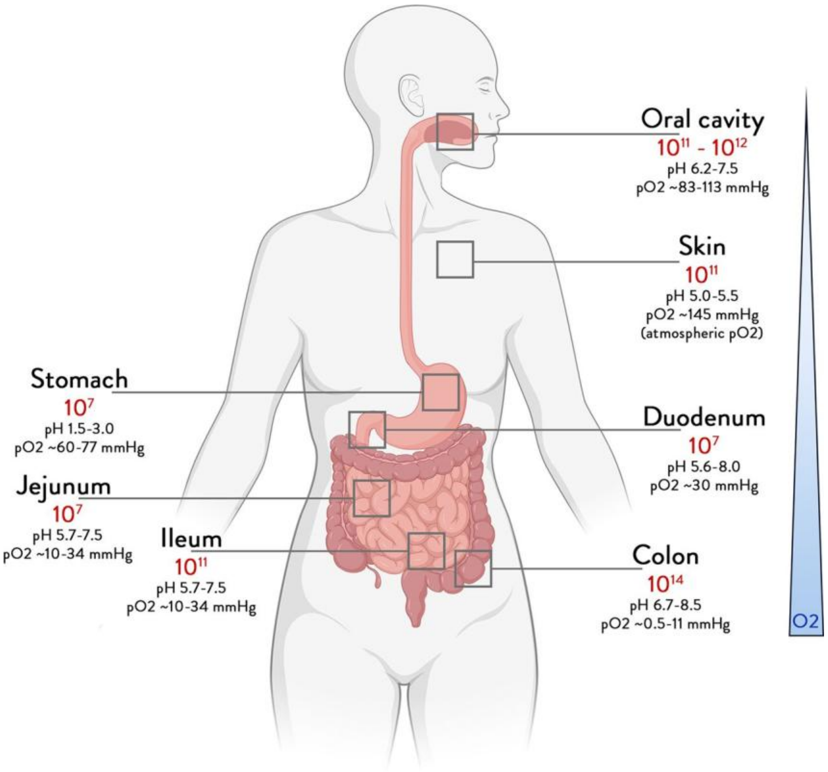


Figure 6_Bacterial Concentrations Based on Regions

Bacterial concentrations differ depending on the body area. The stomach has 10^7 bacterial cells, which is roughly the same as the small intestine (10^7 – 10^{11}). The colon has the highest bacterial load, with 10^{14} cells. The picture is from W. M. de Vos, H. Tilg, M. V. Hul, and P. D. Cani, „Gut microbiome and health: mechanistic insights“[58]

Each of those different areas is not only compiled differently, but they also have different functions. The small intestine is essential to life as it does most food digestion. It also has the highest number of gut receptors, immune cells, and nerve cells that crosstalk with the host [58]. The colon is the space where complex carbohydrates are broken down [55].

1.3.5 Gut Microbiota and Their Influence on Host’s Health

Microbial composition and position influence the host’s health. The microbial influences can be direct or indirect [60]. An indirect influence would be microbial–microbial interaction with the resulting bacterial metabolites. Those metabolites can be small or large molecules whose concentration depends on the abundance of the producing species [58]. Those bacterial metabolites can interact with the host receptors, causing a secondary reaction and leading to the activation or deactivation of metabolic reactions.

Microbiota can alter the metabolism of large molecules such as glycolipids or small metabolites [60]. It also plays a vital role in metabolic reactions such as short-chain fatty acid (SCFA) production, vitamin production, amylolysis, and proteolytic activity. The microbial interference in metabolic reactions can then cause miscommunication for the energy uptake or other signaling pathways. However, their influence on those metabolic pathways follows a daily rhythm which also depends on the food intake [58]. Yet, the influence on the host is major, even if it happens indirectly.

A direct impact would be if the gut function is impaired because of bacterial dysbiosis. If certain bacteria are highly abundant, correlations have been identified to many diseases – internal intestinal diseases as well as external diseases [61]. Direct links have been made to inflammatory bowel disease (IBD), colorectal cancer (CRC), liver disease, pancreatic disorders, but also diabetes or psychological disorders [58], [59].

Inflammatory Bowel Disease

A consequence of microbial disruption is often inflammatory bowel diseases. Scientists found that the gut shows signs of inflammation when dysbiosis and more facultative anaerobes were detected. The microbial disturbance leads to the disruption of the metabolic pathway, which is in charge of short-chain fatty acids production or the acylcarnitine pathway. IBD patients have shown a higher number of *Ruminococcus* [59]. *Ruminococcus gnavus* produces L-rhamnose oligosaccharides, which support the tumor necrosis factor-alpha – a pro-inflammatory cytokine. A permanent overshoot of the cytokine production weakens the immune system and contributes to IBD [59]. Inflammatory bowel disease has two forms: Crohn's disease and Ulcerative Colitis. The difference between them is the area of inflammation in the gut. Patients diagnosed with Crohn's disease have inflamed areas throughout the gut, not limited to a certain area. Whereas Ulcerative Colitis patients have the inflammation confined to the colon mucosa [59]. IBD is characterized by a lot of oxidative stress, leading to decreased microbial diversity. Facultative anaerobes are growing more extensively such as *Enterobacteriaceae* and invasive *Escherichia coli*. Active IBD is often seen with a rapid increase in fungal representation, an increase of lactose fermenting bacteria (*Streptococcus*, *Lactobacillus*, or *Klebsiella*), and human DNA in stool samples from blood or epithelial cells [59]. Crohn's disease also presents with a higher number of *Escherichia coli* and less *Prevotella*. The differences within bacterial abundance and diversity support the physician's diagnosis of IBD, but IBD detection also builds on other markers such as inflammatory proteins, antimicrobial peptides, and SCFA levels. IBD patients have shown that they are low in fecal acetate, propionate, and butyrate but high in lactic and pyruvic acids [59]. Getting diagnosed with IBD is complicated as there is no cure, and current treatment plans build heavily on antibiotics, which only lead to rapid fungal growth but no improvement for the patients. The life quality is low for IBD patients, and in some cases, the disease can progress to life-threatening conditions or cancer [59].

Colorectal Cancer

Colorectal cancer is ranked in 3rd place in cancer diagnoses and ranked 2nd on the cancer mortality scale worldwide. More than two million new cases are added yearly, often in advanced stadiums, as early diagnosis is difficult [62], [63]. Cancer progresses slowly and silently without the patient noticing until it has progressed. In advanced stages, patients complain about abdominal discomfort, alteration in their stool consistency, mucus/blood in stool, and weight loss. In progressed stages, abdominal masses are present as well [62]. Diagnoses are made via the patient's symptoms, physical examination of the rectum and colon via endoscopy, as well as diagnostic imaging such as X-ray, computed tomography (CT), nuclear magnetic resonance (NMR), or positron emission tomography/ computed tomography (PET/CT). Blood work is analyzed to determine tumor markers. However, there are currently no specific markers for CRC [62]. Diagnoses are additionally complex as symptoms can vary. The cancer also diversifies and does not grow in the same places. The large intestine is divided into three major parts: the ascending, transverse, and descending colon [62]. More than half (55 %) of the cancers were found in the sigmoid colon, followed by 23 % in the ascending region. Cancer occurs less often in the transverse colon (8.5 %), descending colon (8.1 %), and the cecum (8.0 %) and is rarely spotted in crossing sites (2.1 %) [62]. The cancer starts with polyps and adenomas, where structural changes in the DNA occur, and normal cells are turned cancerous. Genes such as APC, DCC, P53, k-ras, c-MYC, MCC, and MMRrelated genes are altered. In the early stages, the cancer stays in the submucosa and intestinal mucosa. It is a local invasion. Then, the cancer progresses into malignant tumors, and lymphatic and hematogenous metastasis are spread [62], [64]. In the early stages, patients have the best chance of survival with surgery. Early CRC has a five-year survival rate of 90 % with surgery. Late stages of CRC have a more aggressive take with radiation and chemotherapy but tend to end in passing. To avoid surgery or death, the best treatment for colorectal cancer is prevention. Only 10-35 % of the cases can be blamed on genetics. The remaining cases are all due to environmental factors. External conditions that favor CRC are climate, socioeconomics, education, stress, physical activity, medication, smoking, and diet [63], [64]. Western diet plays a major role in the onset of cancer – high fat, high animal protein, and a limited amount of fresh vegetables and fruits [62], [64]. Food can trigger internal factors for CRC: inflammation, gut microbiome dysbiosis, oxidative stress, and other metabolic consequences [63]. The gut microbiota that colonizes the colon is in constant exchange with epithelial cells, their surrounding microbes, and the host immune system [64], [65]. Studies have shown that *Fusobacterium nucleatum*, *Bacteroides fragilis*, and *Peptostreptococcus. anarobius* promote CRC due to inflammation, bacterial adhesion to host cells, and toxin production. Those pathogens can cause DNA damage and induce pro-inflammatory reactions. *P. anaerobius* activates tumor-promoting pathways, which leads to hyperproliferation of cells. *F. nucleatum* produces an adhesin FadA that starts the signaling pathway for an inflammatory and oncogenic response. Others produce superoxide radicals that damage DNA [64], [65]. Microbiome analyses have revealed that the abundance and composition of microbes on tumor sites differ from the surrounding tissue [65]. Since colorectal cancer patients have shown a higher abundance of *Fusobacterium nucleatum*, *Escherichia coli*, and *Bacteroides fragilis* and are lower in *Firmicutes,* the gut microbiome can potentially be used as a biomarker for diagnostic [56], [58], [64].

Other Diseases Caused by Dysbiosis

Gut microbiome dysbiosis does not only promote IBD and colorectal cancer but does promote other diseases as well*. Akkermansia municiphila* was directly linked to obesity [55], [58]. An increase in obesity was also seen when the plant-degrader *Prevotella* was replaced or outnumbered by *Bacteroides* [59]. It was identified that bacterial LPS introduces inflammation in adipocytes [59]. This has been seen in many people with a high animal protein-based diet, with choline-rich nutrients and a lot of saturated fat. Dietary fiber supports host health as the weight decreases, the blood glucose level stays low, and cholesterol is also low [59]. All of those factors consequently reduce the risk of cardiovascular heart disease and diabetes. Type-2 diabetes has also been correlated with a disrupted microbial composition [55], [58].

A disrupted microbial composition also influences the brain. The brain is connected to the gut, and its communication goes both ways, including the enteric and central nervous systems. If the microbial composition is disrupted or, even worse, a disease like leaky gut syndrome has developed, the communication between the brain and the gut is consequently affected. The leaky gut syndrome causes a weak and porous epithelial wall, allowing bacteria, toxins, and molecules to pass through easily. Consequently, the neuroimmune and neuroendocrine systems are altered, causing changes in brain neurodevelopment [59]. Additionally to the leaky gut syndrome and its association with the brain, scientists have found that bacteria such as *Lactobacillus* and *Bifidobacterium* can metabolize the amino acid glutamate into gamma-aminobutyric acid (GABA). GABA is essential for the GABA receptor signaling pathway and has a direct connection to anxiety and depression [59].

1.3.6 Therapies for Gut Diseases

1.3.6.1 Fecal Microbiota Transplantation

Therapies are limited for all of the above-mentioned gut diseases. Novel attempts to help are fecal microbiota transplantations (FMT). FMT has been shown to improve patients' conditions and fade symptoms. In this therapy, microbes are isolated from a healthy donor and transplanted into the patient’s gut [59], [66]. As described in Figure 7. Despite being relatively new, fecal microbiota transplantation has been around for centuries. The first records are from the fourth century in China, where stool was used against diarrhea. In the 16th century, fermented feces were used to treat diarrhea and abdominal pain [66].

The first medical reports are present from 1958 when scientists successfully treated pseudomembranous colitis. In 2013, the first randomized trial was conducted with patients with recurrent *Clostridium difficile* infection. Stools were transplanted from healthy donors and showed great success, greater than antibiotics on their own. FMT was also successfully tested on Ulcerative Colitis patients in 1989, as the patient showed long-lasting clinical recovery [66]. The FMT success story began when patients with refractory *C. difficile* infections were treated with transplanted stool. If treated with antibiotics, the recovery rate was at 20-30 %; when treated with FMT, 90 % of the patients recovered [66], [67]. Consequently, FMT became widely accepted as a common treatment for *C. difficile* infections, and even the FDA signed off on it in 2013. Scientists also wanted to copy the achievement to other GI diseases but with little to no accomplishment. The more complex the diseases got, the more inconclusive the answers were. One bacterial species is easier to treat than a complex cascade of many unknowns [66], [67]. The fecal donors are selected thoroughly to eliminate as many unknowns as possible. It starts with a health questionnaire and interviews, followed by a medical examination and molecular tests. The optimal donor has no history of GI diseases, rare medication usage, especially antibiotics, no infection of HI virus or Hepatitis virus, does not do drugs, lives in a healthy household, and has no signs of obesity or malnutrition. Since genetics also play a role, non-related donors are often the preferred choice if available [66], [67], [68]. If a suitable donor was found and stool was donated, the stool should be processed rather quickly, e.g., 21 days. For that, the stool is blended with sodium chloride (three times its weight), filtered, and placed in syringes, ready to be administered. [66], [67], [68]. Before administration, patients need to be prepared mentally and psychically: no antibiotics are allowed to be taken, and the GI tract needs to be freed from the stool. After the patient's preparation, fecal matter can be transplanted. Transplantation happens either orally via a capsule or via the upper GI tract nasojejunal, nasogastric, or nasoduodenal via a tube or via colonoscopy in the lower GI [66], [68]. The mode of delivery depends on the patient's willingness and risk management. Capsules are easy, less invasive, and widely accepted but mentally difficult to swallow. Colonoscopy has the advantage that it can examine the area of interest, minimize the risk of residual stool, and place the fecal matter at the target site, but its biggest disadvantage is its invasiveness and costs [66], [68]. Independent of the mode of delivery, risks have to be evaluated individually, along with the side effects that can occur, such as diarrhea, fever, abdominal discomfort, or complications during sedation and endoscopy. Long-term effects are still not fully investigated and can, therefore, not be considered. As with every medical therapy, things can go wrong also in FMT. Some cases have been reported that patients had worse symptoms than before. The exact reasons behind that outcome are unknown. FMT is still in its infancy, and many things are still under investigation. However, the potential of this therapy has been seen worldwide and scientists continue to improve the treatment [66]. During this improvement, a new promising aspect was seen in fecal microbial transplantation when only the viral content was transplanted (FVT). Currently, FMT contains not only bacterial cells but also viral cells, fungi, archaea, metabolites, and eukaryotic host cells [69], [70]. The exact composition in concentration and abundance is unknown and varies a lot based on donors. As a result, reproducibility of results is difficult, if not impossible. Consequently, the idea is to transplant only a single component rather than an entire mix [70]. The main components of stool are bacteria and viral-like particles (phages). Phages have been reported to play a vital role in the human gut, and FMT data has confirmed that phages are more stable in their colonization and remain longer than bacterial transplants [69]. It was also recorded that *C. difficile* patients showed a higher abundance in *Microviridea* and lower levels in *Caudovirales* after FMT. Recurrences or FMT failures have also been associated with different phage compositions compared to successes [70]. First tests have been made with fecal viral transplantation. For that, the stool was blended as previously explained, but before administration, the suspension was filtered via a sterile filter to eliminate all bacterial cells and debris [70]. *C. difficile* patients were symptom-free for six months. Other animal experiments showed promising results [69], [70]. However, since fecal viral transplantation is in its infancy, many more experiments and trials are necessary to totally understand the underlying mechanisms and interactions in the human gut, during diseases, and after FMT/FVT.

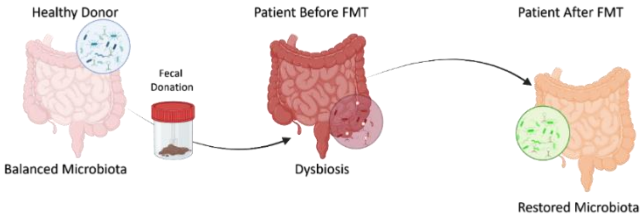


Figure 7_Fecal Microbiota Transplantation Overview

Fecal Microbiota Transplantation is a therapy option for gut diseases. Feces are collected from healthy donors, and microbiota is isolated. The isolated matter is then transplanted into the patient. The transplant can either include the entire microbial diversity – bacteria, viruses, archaea, etc. (FMT) or only the viral content (FVT). The picture was created with BioRender.

1.3.6.2 Probiotics

Probiotics are another approach for gut diseases. Probiotics are bacteria with a proposed health benefit. Probiotics are available over the counter and contain bacteria such as *Bifidobacterium* or *Lactobacillus* spp, which help the gut microbiome [59]. Probiotics support the balance of the gut microbial situation and can boost short-chain fatty acid production. Probiotics have also shown an improvement in acute infectious diarrhea, antibiotic-caused diarrhea, *C. difficile* diarrhea, Ulcerative Colitis, and irritable bowel syndrome. However, therapy with probiotics is not a long-term solution [59]. Probiotics are not licensed medical products. They are not strictly regulated and underly more commercial demands than actual health regulations. The composition and formulation of the product can randomly be changed by the manufacturer, which can lead to full inefficiency. Studies have shown that even bacterial strains from the same genus or species can have a drastically different effect. So, getting the right probiotics for each condition in the right concentration to balance the gut is a challenge [59].

The research continues as the gut microbial challenge is far from being solved. The bacterial influence on gut and host health has been extensively studied and has already shed light onto many dark matters, but now scientists investigate bacteriophages and their role within this whole gut microbial setting -deciphering a new layer of the complex gut matter [57].

1.3.7 Phages and the Gut Microbiota

Bacteriophages have been overlooked for many years, as scientists only focused on bacteria [26]. Bacterial research was more accessible as extensive amounts of bacterial DNA were recovered and characterized based on their 16S rRNA markers. Phage research demanded more. Phages have a small genome size, which makes it difficult to isolate. The little amount of DNA is hard to assemble as phage genomes are firstly without any genetic marker and secondly built like mosaics without structure. Existing assembly tools were quickly exhausted. [26]. Consequently, many new tools were developed to serve the needs of the scientists but without standardization. Standardized protocols are unavailable, and experimental procedures are created on the fly. Not only are wet lab protocols unique, but also bioinformatic processes or databanks are created spontaneously. This lack of standardization in experimental and bioinformatic protocols is prone to user biases. Consequently, results and data interpretation can be difficult and contradictory when studies are compared [61].

Despite these contradictions, scientists found that phages play a vital role in the studied environments as they are the bacteria's natural enemy and can alter the abundance and diversity of the microbial composition. Bacteriophages are primary members of the gut and need to be taken seriously [56].

1.3.8 Phage Distribution

Similar to bacterial distribution, phages are present throughout the body. Their abundance and diversity depend on the location. Phages are abundant throughout the gut, but low levels of phages were determined in the stomach and the beginning of the intestinal tract. Phage abundance tends to increase towards the end of the GI tract [56]. The overall phage-bacteria ratio is around 1:1 in the gut [57], [61]. Phages of the order *Caudovirales* were found throughout and are the majority, followed by *Microviridae* [61].

1.3.9 Phage Colonization in the Gut

Phage colonization starts at the same time as bacterial colonization, as soon as the baby is delivered. Newborn fecal samples are low in phage presence, but phage concentration increases quickly within months [56], [61]. Phageome (entire phage genomes) development is essential but sensitive within the first couple of years [56]. After stabilization, scientists found that healthy individuals have core microorganisms, a core phageome. This core consists of bacteria, lysogens, and free phages [56]. The core can easily be manipulated by food intake – macromolecules, micromolecules, proteins, and carbohydrates. Alterations within the core and gut virome DNA have been linked to gut dysbiosis and disease [56].

1.3.10 Phage Influence on Gut Microbiota

Besides food, the phageome is highly influenced by the host's environment. Geography, culture, and living situations shape the phageome. However, although phages are highly influenced by their surroundings, they are highly influential on their environment as well [56]. Bacteriophages balance bacterial concentrations and alter diversity to keep the equilibrium upright [26]. The interaction is highly dynamic between phages and their hosts and can either be positive or negative for the bacteria [56]. Phages do not only regulate microbial composition; they also carry genes that are influential for anaerobic respiration or biosynthesis of macromolecules, which are transferred via horizontal gene transfer [26]. Horizontal gene transfer from phages to bacteria can increase bacterial pathogenicity via toxin genes. An example is Phi V1/7 phage. Phi V1/7 phage is infecting *Enterococcus faecalis*, and by doing so, it boosts the bacterial pathogenicity and causes more inflammation in the gut [56]. Phages can also add metabolic pathway genes, which can help with polysaccharides and carbohydrates [26]. The Phage CLB_P3 is such an example. It infects *Escherichia coli* strain 55989 and thereby supports biofilm production [56].

In comparison, some phage infections can be beneficial. Prophages are a key part of exchanging genes with bacteria and are responsible for 5 % of nutrient metabolisms and a stable population [56].

1.3.11 Beneficial Relationships Between Phages and Human Individuals

Same as for bacteria and the host, the relationship between phages, bacteria and human host is mutually beneficial. Together, they work to keep the balance and support the host's health.

A health benefit has been seen with an increased abundance of *Caudovirales* and *Siphoviridae* family. Increased levels of those phages were linked to better brain function and verbal memory [56]. Phages also support the immune system by defending the host from pathogens. Phages have surface proteins on their capsid, making them able to bind to the mucus wall and form a layer against pathogens. In addition to that, phages can also attract macrophages and guide them to the intruder [56]. However, phages can also play a major role in gut diseases.

1.3.12 Phages and Gut Diseases

Apart from general functions, phages in the microbiome have also been studied in disease conditions.

Inflammatory Bowel Disease

Inflammatory bowel disease has been associated with many microbial alterations in the gut. Recent findings highlight the role of bacteriophage in IBD. IBD has two forms, and each of those types has unique results. For Ulcerative Colitis patients, a higher level of phage from the *Caudovirales* family was detected but with low diversity. Those patients also had a high concentration of *Escherichia* phages and *Enterobacteria* phages [56], [61]. However, Crohn's disease patients present similar changes as Ulcerative Colitis patients but show an increased level of temperate phages as well. Those virome changes are mirrored in the bacterial composition. Crohn's patients also have more phages attached to the mucus than healthy individuals. Patients with IBD were also diagnosed with lower levels of *Firmicutes* and *Firmicutes* phages [61]. Since IBD is often the starting point for cancer development, scientists also looked into bacteriophages and their role in CRC.

Colorectal Cancer

However, cancer virome studies are scarce. It is known now that bacteriophages are important for the human gut and could be used as a diagnostic tool [71]. To proof that, metagenomic cohort studies were performed, including 317 samples. Five potential phage markers were identified in that study. Those five phages were identified to infect *Fusobacterium nucleatum*, *Peptacetobacter hiranonis*, and *Parvimonas micro* [71]. A second metagenomic study sequenced 462 CRC samples and 449 healthy samples. The study discovered that the alpha diversity was higher in the CRC samples than in the healthy individuals [72]. Additionally, they identified 11 viral species which were enriched (Table 1).

Table 1_List of Enriched Phages

Myovirus	***Podovirus***	***Siphovirus***
Erwinia phage phiEt88	*Salmonella* virus Epsilon15	*Pseudomonas* virus B3
Klebsiella virus ST16OXA48phi5-4,		*Escherichia* phage HK639
Vibrio phage martha 12B12		*Enterobacteria* phage phi80
Mannheimia phage vB_MhM_3927AP2		*Enterobacteria* phage ES18
Salmonella phage 118970_sal3		*Cronobacter* phage phiES15

The study reported that the phage abundance was higher in *Drexlerviridae, Inoviridae, Myoviridae, Podoviridae, and Siphoviridae families;* the only exception was the phage family *Herelleviridae* [72]. For that family the abundance was highly decreased. It was described that these differences in abundance are associated with CRC as well as the presence of *Erwinia* phages and *Vibrio* phages, which are linked to CRC growth [72]. CRC and its virome were also studied under disease conditions, and an animal study showed that temperate phages are highly influential for CRC in the presence of a *Helicobacter pylori* infection [73].

Other Diseases

The influence of phages on other diseases was also investigated. Studies also showed that not only do concentration changes have an influence on the host's health, but ratio alterations can also lead to diseases. If the ratio between *Escherichia coli* phages and *Escherichia coli* cells is disrupted, it can then trigger a prophage induction, which was correlated to Typ-1-diabetes [56].

Patients with irritable bowel syndrome showed a higher degree of similarity of virome compared to healthy individuals and showed a higher degree of diversity within each individual [61].

Patients with metabolic syndrome showed increased levels of phages infecting *Streptococcoceae* and *Bacteroidaceae* and drastically low levels of phage for *Bifidobactericeae*. The overall diversity and richness were different compared to healthy individuals [56].

1.3.13 Objectives

Many things have been discovered already, but the majority is still unknown. How phages, bacteria, and the remaining gut microbiota interact with each other and, by doing so, influence the host's health. In the last chapter of my thesis, we aim to analyze the role of phages in healthy individuals, patients with Ulcerative Colitis, patients with early colorectal cancer and in patients with advanced colorectal cancer.

2. Material and Methods

Viral tagging has been developed by Deng et al [74] but has been optimized for human pathogens and various sorting machines by myself for this thesis. This thesis contains three projects, all with viral tagging as their main method. However, despite having viral tagging as their main method, each project has a slightly different variant based on the project aim. In the next chapter, the workflow of each individual project is described, including all details and method variants. Additionally, all materials used for this thesis are listed at 2.4 Material

2.1 Targeted Single Phage Isolation

This method has been published by Unterer et al. and pictures were taken from the publication [75].

The general workflow is shown in Figure 8.

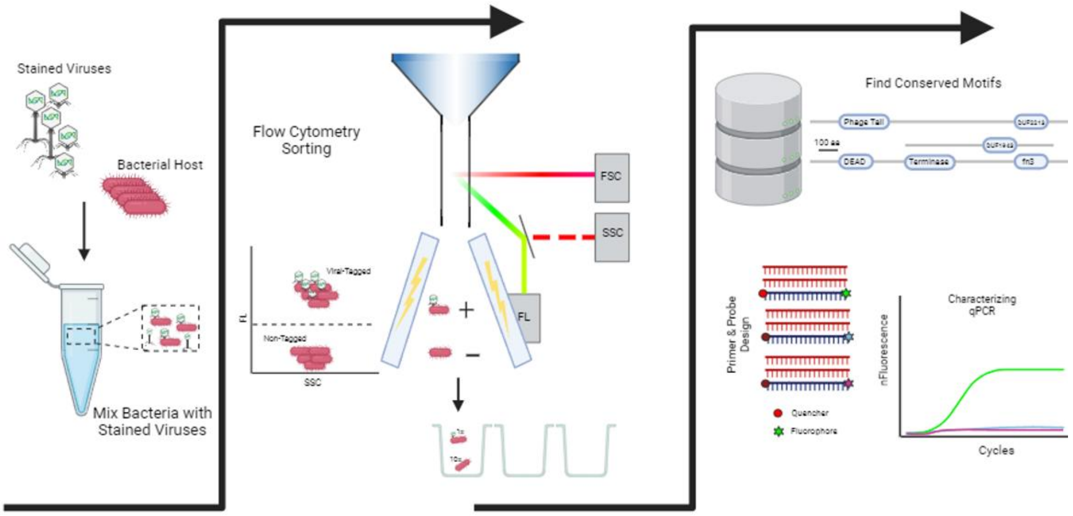


Figure 8_General Workflow for Targeted Single Phage Isolation

Stained virus-like particles are mixed with bacteria before viral-tagged cells are separated from non-tagged cells via flow cytometry. Single-tagged cells are then characterized via infection dynamics and qPCR. Published in Unterer et al. [75]

2.1.1 Bacteria and Phage Cultivation

“Targeted Single Phage Isolation” was developed using the bacterium *Escherichia coli* 11303 and its three model phages, T1, T4, and T7. Those three phages were chosen because they have been studied and used widely.

The strain was cultivated on Luria Bertani Broth (LB, Carl Roth, X968.1) agar plates and LB liquid medium at 37 °C aerobically. For every experiment, fresh cultures were prepared.

At the beginning of the project, fresh phage stocks were produced. To get those, bacteria overnight cultures were diluted and incubated until cells reached the stationary phase. At this stage, phage stock was added, and the mixture was incubated overnight on a shaker. On the next day, the culture was centrifuged at 5,000 rcf for 15 minutes, and the supernatant was filtered with a 0.22 µm syringe filter (Merck, Millex-GP, SLGP033RS). The new phage stock was put into the fridge for further usage. Each phage was grown up separately.

2.1.2 Phage and Virus-Like-Particle Counting

Phage experiments can only be performed if the amount of phage particles in the solution is known. Double-layer plaque assays were carried out to determine the amount of phage particles in the stocks. In detail, phage stocks were diluted with phosphate-buffered saline (PBS) to reach a dilution of 10^{-3} to 10^{-9}. Then, 3 mL LB soft agar (0.6 %) was mixed with 200 µL bacterial culture and 100 µL phage dilution and poured onto an LB agar plate (1.5 %). Plates were incubated upside-down at 37 °C overnight. On the next day, plaque-forming units(PFU) were counted, and PFU/mL was calculated with the formula below.

$$\frac{PFU}{mL} = \frac{Number\ of\ Plaques}{Dilution\ Factor\ x\ Volume\ of\ Phages\ added\ (mL)}$$

Phage titers can only be determined via plaque assays if bacteria and phages are cultivatable. For unknown sources, only the amount of virus-like-particles (VLP) can be detected. To count the VLPs in the high-speed concentrated samples, I used the Nanoparticle Tracking Analysis device (NTA, NanoSight, Malvern Panalytical). VLPs were stained with 1 µL SybrGold (50x SybrGold) for 30 minutes at 30 °C. Stained samples were diluted 1:100 or 1:1000 with PBS and processed with the analyzer. Based on Brownian motion and fluorescence, the amount of virus-like particles can be calculated. The final concentration was determined by the machine based on the dilution factor.

2.1.3 Single Phage Viral Tagging and Infection Dynamics

For the first part of the method, phages were stained with Syto9 (ThermoFisher, S34854) for 30 minutes in the dark at room temperature. After incubation, they were washed three times with water (Milli-Q, Merck) using Vivaspin sample concentrator (Vivaspin 20, 100,000 MWCO PES, VS2042) at 3,000 rcf for 3 minutes. After the washing step, stained phages were kept on ice.

The bacterial overnight culture was washed thrice at 5,000 rcf for 3 minutes with saline solution and kept on ice.

Washed bacteria were mixed with stained phages at a ratio of 1:2 and incubated in a thermocycler at 37 °C for five minutes. Before the sort, the mixture was diluted 1:1000 with water.

Single-cell sorting was performed on b.sight (cytena). Before the experiments could start, a quality control step was executed as the company suggested. For that, the camera was aligned, the droplet was centered, and the cartridge was visually checked for any debris or bubbles. After a successful quality check, the cartridge was filled with the sample, and sorting settings were adjusted (Cell Size: 0.8 to 3, Cell Surface: 0.5 to 1, Fluorescence Intensity: 46 to 56, Fluorescence Size: 0 to 10). Ten bacterial cells were sorted into each well (filled with 200 µL LB), before one viral-tagged cell was added. Bacterial cells-only control had only ten bacterial cells inside the wells, whereas media-only had no cells at all.

To determine the infection dynamics of sorted phages, cells were incubated shakenly in a plate reader (BioTek EPOCH2) at 37 °C for at least 18 hours. Optical density at 600 nm was measured every 20 minutes.

2.1.4 Phage Characterization

For further characterization, a multiplex qPCR was developed which can identify multiple targets within one reaction. Primers and probes were designed based on conserved motifs from chosen targets [76] – capsid protein for T1, T4, and T7, eae/intimin, and LT toxin (heat-labile toxin) (Supplementary Table 1), which was created by PrimerQuest Tool from Integrated DNA Technologies. Probes were equipped with Cy5, FAM, or HEX fluorophores, which were quenched by BBQ650 or BHQ1, respectively (Supplementary Table 2). To avoid hairpin formation or dimers, primers, and probes were tested via DINAMelt Server. qPCR reactions were performed on Agilent

Mx3000P at a total volume of 20 µL containing three primers sets (300 nM each), three probes (100 nM each), reference dye (ROX), 1x master mix (Agilent, 600553) and water. The total volume also included either 1 µL of the sample (1:10 diluted), 1 µL of lambda DNA as a negative control, or 1 µL water as no template control. After a ten-minute initial denaturation phase, 40 cycles followed with 15 seconds at 95 °C and 1 minute at 62 °C. Data were normalized by the machine based on its internal control. Additionally to the internal control, DNA standards for each primer set were added onto each plate ranging between 10 ng to 0.1 pg. The data was visualized with a GraphPad prism.

Phage DNA used for the standards was extracted by the Norgen Phage DNA Isolation Kit (Norgen Biotek Corp., Cat. 102 #46800) as the manual instructed and bacterial DNA was isolated with the DNeasy PowerLyzer Microbial Kit (Quiagen, Cat. No. / ID: 12255-50) as said in the manual.

2.1.5 Method Validation Tests

To test if Syto9 has a negative effect on phage infectivity, one aliquot of phages was stained with 1 µL Syto9 for 30 minutes at room temperature in the dark, whereas the second aliquot was incubated at the same condition but without dye. After incubation, both aliquots were washed three times with water in an ultrafiltration unit (Vivaspin 20, 100,000 MWCO PES, VS2042) at 3,000 rcf for three minutes. Washed phages were then tested in double-layer plaque assays at different concentrations to determine the infectivity after the process.

Additionally to the cytotoxicity testing, I tested if the developed method had an effect on the phage. Therefore, I did a one-step growth curve to get comparable data. Phages were prepared as usual (protocol above) and mixed with bacteria to reach a multiplicity of infection (MOI) of 0.01. 200 µL of that mixture was added to a 96-well plate. The plate was incubated at 37 °C in the plate reader with continuous shaking for one hour. Every five minutes, 100 µL samples were taken, mixed with 900 µL LB, and filtered through 0.22 µm to eliminate all bacterial cells. The filtrate was then again diluted, and phage titer was determined via double-layer plaque assay. The experiment was done in triplicates for all three phages.

To further validate the specificity of the method, wastewater was spiked with model phages at three different concentrations: 1,000x more phages than wastewater VLPs, at equal amounts, and 1,000x fewer phages than wastewater VLPs. These mixtures were tested on a single-cell level, sorted with the single-cell dispenser into a 96-well plate, and on a flow cytometer with 1 million sorted cells. The protocol for single cells was the same as explained above. For 1 million cells, the protocol was adapted for the flow cytometer, MoFlo (Beckman Coulter, USA) since additional controls were needed. For that, bacterial overnight cultures were washed three times with saline solution for three minutes at 5,000 rcf. One aliquot was stained with 1 µL Syto9 for 30 minutes at room temperature in the dark. After a second round of washings, bacterial controls were stored on ice. Buffer controls were handled the same way as VLPs. Before viral-tagged cells could be sorted, the flow cytometer needed a quality control run. Lasers and stream were aligned with beads (CytoFLEX Daily QC Fluorospheres, Beckmann Coulter, B53230), drop delay adjusted, IntelliSort activated, and tubes were cleaned. After everything was ready to sort, samples were processed in the following order: unstained buffer, unstained VLPs, unstained bacteria, unstained bacteria with unstained VLPs, water, stained buffer, stained VLPs, stained bacteria, water, unstained bacteria with stained VLPs. Samples were recorded with a slow flow rate until 10,000 events or 10 seconds were reached. Cells were visualized on a logarithmic scale on Side Scatter vs Forward Scatter plots (FSC) and Fluorescence vs Side Scatter plots (SSC). Stained and unstained bacteria were gated to determine the viral-tagged cells. Data files (.fcs) were analyzed with FlowJo (10.8.1 CL).

Single sorted cells and one million sorted cells were then detected via qPCR as explained previously.

2.2 *Helicobacter pylori* Phages and Their Diversity

Although bacteriophages are highly abundant, isolating them is challenging due to their host specificity. Therefore, we have chosen sources where phages and hosts co-exist to increase the probability. And since *Helicobacter pylori* colonizes the gut, we took feces and sewage water. The overall workflow is shown in the Figure 9.

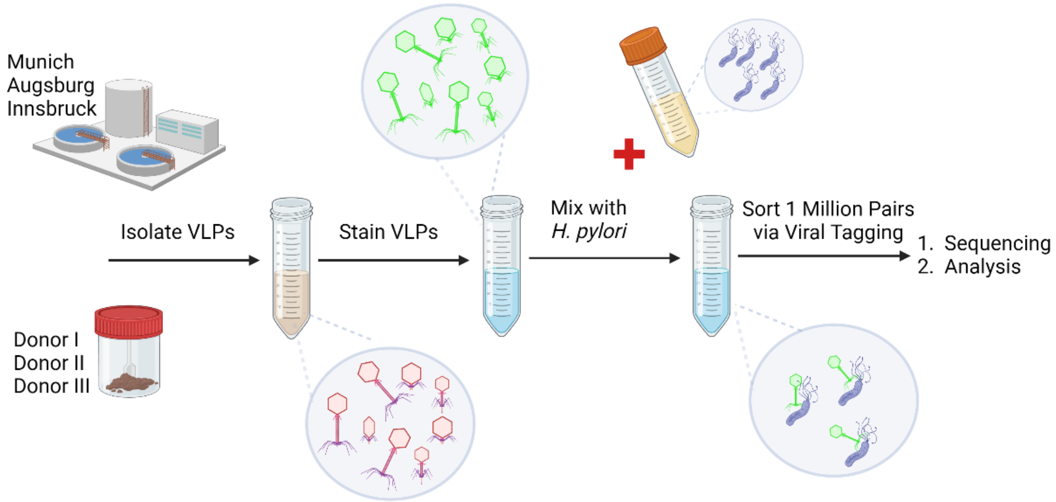


Figure 9_General Workflow for Isolating *Helicobacter pylori* Phages

Virus-like particles were isolated from wastewater and fecal samples, stained, and mixed with *Helicobacter pylori*. Viral-tagged cells were isolated via flow cytometry and sequenced. The picture was created with BioRender.

2.2.1 Sample Collection

To increase the chances of phage isolation, I used wastewater and feces for this study. 500 mL active sludge (active microbial content [77]) were collected from three different wastewater plants, Munich and Augsburg (Germany) and Innsbruck (Austria), and transferred directly to the lab, where the water was stored at 4 °C until processed. Human stool samples were donated and, upon arrival, directly stored at -80 °C.

2.2.2 Bacteria Cultivation

Two closely related *Helicobacter pylori* strains PMSS1 and SS1 were cordially provided by Prof. Dr. Markus Gerhard (Technical University of Munich). Strains were grown on blood plates supplemented with *Helicobacter pylori* Selectives (M863-500G, HIMEDIA+ horse blood, SR0147E, Oxoid) at 37 °C in a microaerophilic environment (10 % CO_2, 5 % O_2). For each experiment, a bacterial overnight culture was set up in 15 mL liquid media containing BHI, 20 % FCS and supplements (X916.2, ROTH, SR0147E, Oxoid). On the day of the experiment, the quality of liquid cultures was tested via urase broth test (M1828-500g, HIMEDIA) and microscopy.

2.2.3 Isolation of Phages From Feces and Wastewater

The stool sample was prepared fresh for each experiment. The stool was weighted and mixed with PBS at a ratio of 1:10 (w:v). The mixture was vortexed thoroughly for one hour. After the mixing, the samples were centrifuged at 700 rcf for one minute to remove fecal matter and big debris. The supernatant was then again centrifuged at 6,000 rcf for 45 minutes to separate bacteria and VLPs. The new supernatant was filtered through 0.22 µm syringe filter (SLGP033RS, Millex-GP Syringe Filter, Merck) before it got concentrated via an ultrafiltration unit (Vivaspin 20, 100,000 MWCO PES, VS2042, Sartorius) to 1 mL. The concentrate was stored on ice.

Since active sludge does not only contain active microbial cells but also has big debris and particles, wastewater was centrifuged at 6,000 rcf for 15 minutes before filtering. The supernatant was filtered with a 0.22 µm syringe filter (SLGP033RS, Millex-GP Syringe Filter, Merck) before 500 mL filtrate was concentrated via high-speed centrifugation (35,000 rcf for 2 hours). The pellet was resuspended in 5 mL PBS and stored at 4 °C.

2.2.4 Bacteriophage Isolation via Viral Tagging

To separate *H. pylori-specific* phages from all the others, I performed the previously established viral tagging technique. For that, bacteria, buffers, and virus-like particles in sufficient amounts were needed.

The bacterial overnight culture was transferred into two reaction tubes where they were washed three times with saline solution (0.9 % NaCl) at 5,000 rcf for three minutes. After the washing, one aliquot was stained with Syto9 (1 µL, ThermoFisher, S34854) for one hour at room temperature in the dark. To remove the residual dye, a second round of washings as before was performed but for both aliquots. Until further, bacterial samples were stored on ice.

As a next step, buffer and VLP samples were prepared. For both, two aliquots were pipetted, and one of each was stained with 1 µL of Syto9 (ThermoFisher, S34854). All four tubes were incubated at room temperature for 30 minutes in the dark. Before, they were washed three times at 3,000 rcf for 3 minutes with water in an ultrafiltration unit (Vivaspin 20, 100,000 MWCO PES, VS2042, Sartorius). To store them until the run, they were put on ice.

For the viral tagging samples, 100 µL unstained bacteria were mixed with either 200 µL stained or unstained VLPs. Mixed and incubated in a thermocycler for one hour at 37 °C. To ensure a smooth sample recording, bacterial samples were diluted 1:1000, VLPs 1:100, viral-tagged samples (VTs) 1:100, and buffer samples were recorded undiluted. All samples were processed with a flow cytometer BDFACS Melody (BD), and one million cells were sorted into 200 µL PBS and stored at 4 °C.

Before I could process all samples, a quality control run was performed as the company suggested. The lasers were aligned, and the correct drop was established. After a successful QC run, the machine was cleaned to eliminate the remaining QC beads and avoid any form of cross-contamination. To minimize the risk of spillovers or contaminations, samples were processed in the following order: unstained bacteria, stained bacteria, water, unstained buffer, unstained VLPs, unstained bacteria with unstained VLPs, water, stained buffer, stained VLPS, water, unstained bacteria with stained VLPs. Samples were processed at a slow flow rate and also recorded at that speed. I recorded either 10,000 cells or 10 seconds. For visualization, I used two plots, Side Scatter vs Forward Scatter and Fluorescence vs Side Scatter, to create our gates. Plots were analyzed with FlowJo v10.8.1. (FlowJo, BD)

2.2.5 DNA Extraction

Single Bacteria

DNA from *Helicobacter pylori* strain PMSS1 and SS1 was extracted using the DNeasy PowerLyzer Micrbioal kit (12255-50, Qiagen) according to the company's instructions. Extracted DNA was concentrated and purified using the Genomic DNA Clean & Concentrator kit (D4064, Zymo Research) as instructed by the manual. DNA concentration was measured via Qubit. DNA was stored at -20 °C until it got sent for sequencing.

Virome and Sorted Cells

DNA from wastewater VLPs, stool VLPs, and sorted cells were isolated identically with the following protocol: the sample solution was concentrated to 500 µL with an ultrafiltration unit (Vivaspin 20, 100,000 MWCO PES, VS2042, Sartorius) before external DNA got degraded via DNAse I (EN0521, thermoscientific) for 30 minutes at 37 °C. To inactive the enzyme, samples were incubated for ten minutes at 70 °C and then placed at -80 °C for one hour. The physical disruption continued with a defrosting step at 65 °C for 30 minutes. To eliminate remaining proteins, samples were treated with proteinase K (20 mg/mL, AM2548, invitrogen) for 30 minutes at 55 °C and the enzyme got inactivated afterward at 70 °C for 15 minutes. After all enzymatic and physical treatments, samples were mixed at a ratio of 1:2 with AMPure XP beads (A63881, Beckman Coulter) and incubated at room temperature for 15 minutes. Two wash steps were performed to wash out every impurity with 70 % ethanol. After the entire ethanol was evaporated, beads were incubated with elution buffer for five minutes at room temperature for final DNA extraction. To gain ultra-pure and concentrated DNA, every sample was processed through the Genomic DNA Clean & Concentrator kit (D4064, Zymo Research).

Since the DNA concentration of those samples was too low, DNA got amplified by the REPLI-G kit (150343, Qiagen) for single cells (as recommended by the company). Amplified DNA was measured via Qubit and sent for sequencing.

2.2.6 Sequencing

The Illumina platform Novaseq was used to sequence the sample with their library prep, TruSeq. 2 gigabytes (GB) of data were acquired for each sample via a 2x150 paired-end (PE) strategy. The sequencing was outsourced to a sequencing facility.

2.2.7 Data Analysis

Helicobacter pylori

Reads quality was controlled by fastp [78] (v0.23.2) to then assemble the clean reads with SPAdes (v3.15.2) [79]. After de-novo assemble, we conducted a reference-based assemble process to merge the contigs to draft genomes by Rebaler (v0.2.0; https://github.com/rrwick/Rebaler) which relied mainly on minimap2 for alignment and Racon for making consensus sequences. The reference genomes used in Rebaler were the following: for *Helicobacter pylori* PMSS1, they were GCF_001991095.1 and GCF_004295545.1, and for Helicobacter pylori SS1, it was GCF_002005525.1. After that, the clean reads were used to polish the draft genome using Pilon (v1.24) [80]. We used Prokka (v1.14.5) [81] to annotate these draft bacteria genomes with parameter "--evalue 1e-05 --coverage 50 --gcode 11 --kingdom Bacteria". Genome maps were visualized by Proksee (web version) [82].

In-House *Helicobacter pylori* Database

To better refine *Helicobacter pylori* phages, we first created a *H. pylori* phage database which contained public *H. pylori* phages and *H. pylori* prophages. The database was created the following: we downloaded (a) 41 *Helicobacter pylori* phages from the National Center for Biotechnology Information (NCBI) RefSeq [83]; and then (b) 389 *Helicobacter pylori* bacteria sequences whose assemble level was equal to "Chromosome" or "Complete Genome". We predicted 90 prophage regions from these *Helicobacter pylori* bacterial genomes using PhiSpy (v4.2.21, use vog213 database) and Phigaro (v2.3.0). In total, the in-house databank combined 90 predicted prophages and 41 existing *Helicobacter pylori* phages.

Viral Tagging Reads

We first used fastp [78] (v0.23.2) to control the reads quality, then the clean reads were assembled using SPAdes (--meta, v3.15.2)[79]. Contigs with lengths longer than 1,000 bp were kept for further analysis. Then CheckV (v0.8.1) [84] was used to remove the host region and assess the contig's quality. We also applied VirSorter (v1.0.6) [85] to identify the viral contigs as a CheckV complementary. We picked viral contigs if one of the software (CheckV or VirSorter) predicted the contig as a virus and the length was longer than 10 kbp. Host information of all the potential viral contigs was predicted with

Integrated Phage-Host Prediction (iPHoP)(v1.1.0) [86]. As not all contigs could be predicted by iPHoP, we also aligned these potential viral contigs to the Nucleotide Database (NT) prokaryotic and NT virus database separately via BLASTn (v2.13.0) (download 2023-07-26) to get more host information. Additionally, to ensure that we do not exclude any prophages during the alignment to the NT prokaryotic sequences, we also used BLASTn (v2.13.0) with parameter "-max_target_seqs 5 -qcov_hsp_perc 50" in our in-house *Helicobacter pylori* database and 389 public *H. pylori* bacterial sequences.

We categorized all the potential viral contigs into three categories based on the results from above:

Category 1 (Cat 1) included all viral contigs with *Helicobacter pylori* as their predicted host. The prediction was done by iPHoP.

Category 2 (Cat 2) had all novel potential *Helicobacter pylori* phages which had no host prediction by iPHoP and could not be mapped to NT prokaryotic nor NT viruses database.

Category 3 (Cat 3) were all phages that were found in our in-house *Helicobacter pylori* database (phage and prophage).

To further investigate these *H. pylori* phages identified from viral tagging samples, we built a pseudo-phylogenetic tree using pairwise Average Nucleotide Identity (ANI) value with the in-house *H. pylori* database. First, we used CDHIT (v4.8.1, psi-cd-hit, -c 0.90 -G 1 -g 1 -aL 0.7 -aS 0.7 -prog blastn -circle 1) [87] to remove redundancy of potential *Helicobacter pylori* phages. 491 non-redundancy potential phages were retained. To facilitate rerooting the tree, we used SARS-CoV-2 (NC_045512.2) and three *Helicobacter pylori* bacterial sequences as outgroups. These three bacterial sequences were from PMSS1 (GCF_001991095.1; GCF_004295545.1) and SS1 (GCF_002005525.1), which were the hosts used during the viral tagging process. Then hierarchical clustering was built based on pairwise ANI values calculated using FastANI (v1.33).

There were four highly potential *Helicobacter pylori* contigs from Cat 1 whose lengths were longer than 10 kbp and the genome quality was above the median quality assessed by CheckV [84]. These genomes were further investigated by using Prokka to annotate (parameter "--evalue 1e-05 --coverage 50 --gcode 11 --kingdom Bacteria") (v1.14.5) [81] and were visualized by Proksee (web version) [82]. We also checked 34 contigs from Cat 2 with lengths greater than 10 kbp, but they could not be mapped to any genomes from the NT prokaryotic or viruses database. We first used CDHIT (v4.8.1, psi-cd-hit, -c

0.60 -G 1 -g 1 -aS 0.6 -prog blastn -circle 1) [87] to remove redundancy, and in total 25 contigs were retained. The comparison genomic map was created using Clinker (v0.0.23) [88].

Auxiliary metabolic genes (AMG) genes were identified in these novel HP contigs using DRAM-v (v) [89] with parameter '--skip_trnascan --min_contig_size 1000 --prodigal_mode meta --trans_table 11'

Virome Data

Reads were prepared as previously described above but then we used methods established by Nayfach et.al [84]. In short, BLASTn was used to calculate pairwise ANI value by combining all-vs-all alignments between sequence pairs. Then, UCLUST-like clustering was done using the MIUVIG recommended parameters (95 % ANI + 85 % AF). After we got the non-redundant virome dataset, contigs from viral tagging were aligned to contigs assembled from virome using BLASTn (v2.13.0). We also created a Principal Component Analysis (PCA) plot using the ANI value to compare the genome sequences identified by virome and viral tagging.

Functional Annotation and Endolysin Tree

Functional comparison was done with all 599 viral-tagged contigs plus all sequences from the in-house database. Proteins were first predicted using Prodigal (v2.6.3)[90] and then annotated against four databases: Pfam v34 [91], KEGG (download date: 2022-02-01) [92], VOGDB (211, https://vogdb.csb.univie.ac.at/), PHROGs (http://millardlab.org/2021/11/21/phage-annotation-with-phrogs/) [93] under 1e-5 e-value criteria using HMMER (v3.3.2) [94]. MMseqs2 easy-cluster command (version: 7aade9df7475ae7c699b2074b5e4daa52e0245f1; --cov-mode 0 -c 0.70) [95] was used to cluster these proteins. The function of the cluster is based on the most frequently detected annotation.

For the endolysin phylogenetic tree, all determined endolysins were used. The amino acid sequences of these endolysin genes were aligned using MAFFT (v7.505, mode: mafft-linsi) [96] and gaps were removed using trimAl (v1.4.rev15l, -gappyout) [97]. The tree was computed using IQ-TREE (v2.0.3) [98] with 1000 ultrafast bootstrap replications and the VT+G4 substitution model, as suggested by ModelFinder. The tree was visualized using iTOL [99].

2.3 Colorectal Cancer Cross-Infection Study

To better understand the role of phages within disease conditions, single-cell cross-infections were performed between bacteria and virus-like particles from four conditions: healthy donors, Ulcerative Colitis patients (UC), early cancer (CRCE), and advanced cancer patients (CRCA). The general workflow is shown in Figure 10.

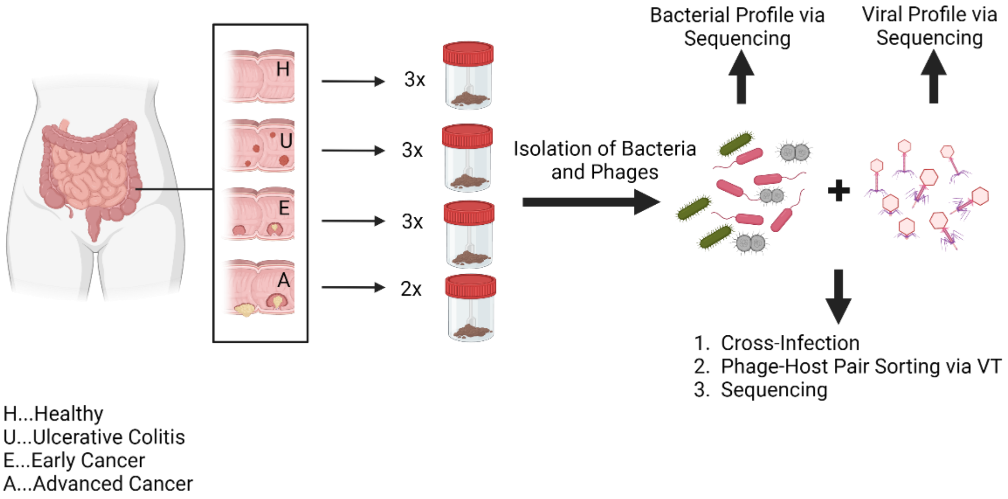

Bacterial Profile via Sequencing
Viral Profile via Sequencing
H
U
E
A
3x
3x
3x
2x
Isolation of Bacteria and Phages
1. Cross-Infection
2. Phage-Host Pair Sorting via VT
3. Sequencing
H...Healthy
U...Ulcerative Colitis
E...Early Cancer
A...Advanced Cancer

Figure 10_General Workflow for the CRC Cross-Infection Study

Bacteria and virus-like particles were isolated from patient samples. VLPs were stained and cross-infected with bacterial samples from all four conditions. 100 viral-tagged cells were isolated and sent for sequencing. The picture was created with BioRender

2.3.1 Sample Collection

To determine phage-host networks within healthy and disease conditions, stool samples were collected from healthy donors as well as from patients with Ulcerative Colitis, early-stage colorectal cancer, and late-stage colorectal cancer at Klinikum Rechts der Isar (Munich, Germany) by the group of Prof. Dr. Klaus-Peter Janssen. After the collection, samples were stored at -80 °C.

2.3.2 Isolation of Bacteria and Phages From Stool Samples

The entire bacteria and phage isolation was done anaerobically for patient stool samples. Stool samples were transferred into the anaerobic chamber, weighed and at a 1:10 ratio (w:v) mixed with reduced PBS. Vortexed thoroughly and centrifuged at 700 rcf for one minute to separate chunky fecal matter from the rest. The remaining sample was then centrifuged again at 6,000 rcf for 45 minutes. At this step, virus-like particles (VLPs) were separated from bacterial cells. The supernatant was then filtered through a 0.45 µm syringe filter (SLHV033RS, Merck) and separated into two piles: DNA extraction and pooling to do cross-infection. The same procedure was repeated for bacterial cells, one aliquot was set aside for DNA extraction, the other one for cross-infection pooling. Samples for DNA extraction were frozen away, whereas the others were kept in the anaerobic chamber.

2.3.3 Cross-Infection and Viral Tagging

All steps were done anaerobically, apart from VLP preparation and sorting.

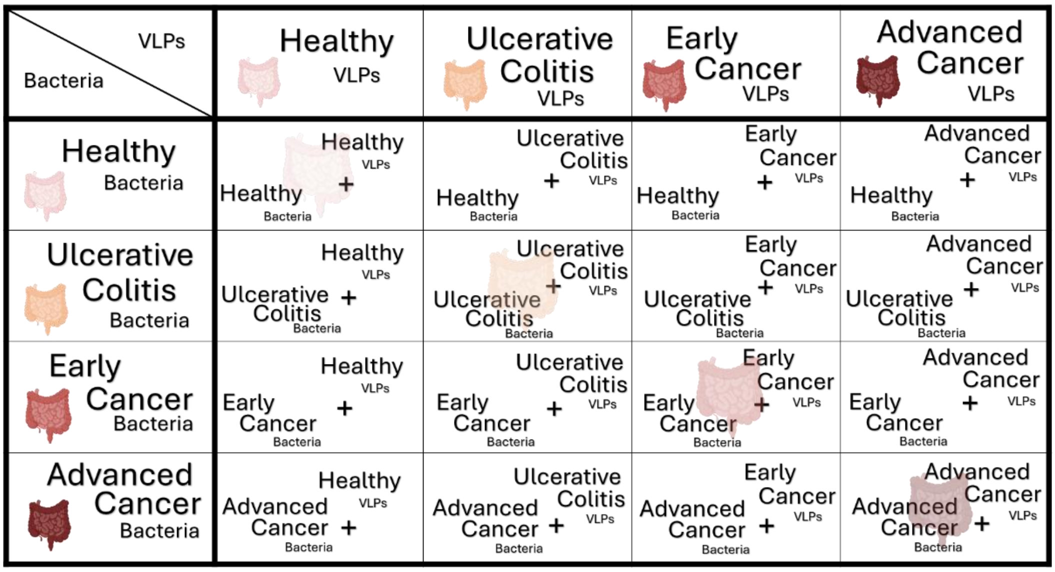
Viral Tagging

Bacterial samples from the same conditions were pooled. 1 mL was taken and washed three times with reduced PBS at 5,000 rcf for three minutes and set aside.

The same pooling step has happened to the VLPs. Pooled samples were then stained with 1 µL Syto9 (ThermoFisher, S34854) for 30 minutes at room temperature in the dark. After the incubation, samples were washed three times with water (Milli-Q, Merck) in an ultrafiltration unit (Vivaspin 20, 100,000 MWCO PES, VS2042) for three minutes at 3,000 rcf to eliminate the remaining dye. After VLPs were stained, all samples were transferred back into the anaerobic chamber.

Cross-Infection

To determine phage-host pairs in each condition as well as their network in diseases, bacterial cells were mixed with different VLPs. 100 µL of bacterial cells were mixed with 200 µL of stained VLPs and incubated at 37 °C anaerobically in the dark on a shaker for one hour. In total, I had 16 variations, which are schematically visualized in Figure 11.



VLPs / Bacteria	Healthy VLPs	Ulcerative Colitis VLPs	Early Cancer VLPs	Advanced Cancer VLPs
Healthy Bacteria	Healthy Bacteria + Healthy VLPs	Healthy Bacteria + Ulcerative Colitis VLPs	Healthy Bacteria + Early Cancer VLPs	Healthy Bacteria + Advanced Cancer VLPs
Ulcerative Colitis Bacteria	Ulcerative Colitis Bacteria + Healthy VLPs	Ulcerative Colitis Bacteria + Ulcerative Colitis VLPs	Ulcerative Colitis Bacteria + Early Cancer VLPs	Ulcerative Colitis Bacteria + Advanced Cancer VLPs
Early Cancer Bacteria	Early Cancer Bacteria + Healthy VLPs	Early Cancer Bacteria + Ulcerative Colitis VLPs	Early Cancer Bacteria + Early Cancer VLPs	Early Cancer Bacteria + Advanced Cancer VLPs
Advanced Cancer Bacteria	Advanced Cancer Bacteria + Healthy VLPs	Advanced Cancer Bacteria + Ulcerative Colitis VLPs	Advanced Cancer Bacteria + Early Cancer VLPs	Advanced Cancer Bacteria + Advanced Cancer VLPs

Figure 11_Schematic for Cross-Infection Study

This schematic shows how samples were cross-infected.

Cell Sorting

To determine those networks, 100 cells were sorted on a single-cell dispenser (bf.sight, Cytena). Before each experiment, a quality control check was performed to align the cameras, to check the position of the droplet and to ensure a good working condition of the cartridge. Sort settings were adjusted between 0.8 to 3 for Cell Size, 0.5 to 1 for Cell Granularity and 46 to 56 for Fluorescence Intensity, and 0 to 10 for Fluorescence Size.

2.3.4 DNA Extraction

Bacteriome

To extract DNA from mixed bacterial suspensions, samples were heated to 70 °C for 15 minutes before I used the E.Z.N.A Soil DNA Kit from Omega (D5625-02, Omega) according to the manufacturer's instructions.

Virome

For virome DNA, samples were concentrated with ultra-centrifugal units (UFC910024, Merck) before they were treated with DNase I (EN0521, thermoscientific) for 30 minutes at 37 °C. For deactivation of the enzyme, the samples were incubated for 15 minutes at 70 °C before they were placed into a –20 °C freezer overnight. On the next day, samples were defrosted at 65 °C for 30 minutes with a subsequent Proteinase K treatment. (55 °C, 30 minutes, 20 mg/mL, AM2548, invitrogen). The enzyme was deactivated at 70 °C for ten minutes. To get hold of the DNA, samples were mixed with AMPure XP beads (A63881, Beckman Coulter) (1:2 ratio) and incubated at room temperature for 15 minutes. Samples were washed two times with 70 % ethanol before incubating them for five minutes with elution buffer to free the DNA. Extracted DNA was purified and concentrated with the Genomic DNA Clean & Concentrator Kit (D4064, Zymo Research) and stored afterward at –20 °C.

Sorted Cells

Sorted cells were used as they were without DNA extraction step.

2.3.5 DNA Amplification

Sorted cells, extracted virome, and bacteriome DNA were amplified with the REPLI-G kit (150343, Qiagen) for single-cells according to the manufacturer's instructions. Amplified DNA was sent for sequencing after a qubit measurement.

2.3.6 Sequencing

Samples were sequenced on the Illumina platform Novaseq with the TruSeq library. For each sample 2 GB of data was generated with the paired-end 2x150 strategy.

2.3.7 Analysis

Native Fecal Metagenome

Raw reads from fecal metagenomes (11 samples in total) were cleaned from PhiX (reference: NC_001422.1) and human (reference: GRCh38) contamination using Bowtie2 (v2.3.5.1) [100] with parameter "--sensitive-local" and SAMtools (v 1.17) [101]. Then low-quality reads were deleted with fastp (v0.23.2) [78] on the following parameter "-z 4 -n 10 -l 60 -5 -3 -W 4 -M 20 -c -g -x". Clean reads were assembled from each sample individually using metaSPAdes (v3.15.2) [79] with the default setting. Scaffold lengths were kept when they were longer than 1 kbp for further analysis. A public human gut database was used as a reference to complement the de novo scaffolding to prevent the loss of short scaffolds. Clean reads were mapped from each sample to the Unified Human Gastrointestinal Genome (UHGG) catalog (v2.0.2) [102] using Bowtie2 (v2.3.5.1) [100] with parameter "--sensitive-local". Sequences were retained if they met the following criteria: at least 80 % of maximum mapping reads were met, or sequences had over 60 % mapping coverage, or their mapping depth exceeded 80 % of the maximum depth based on SAMtools (v 1.17) [101] outputs. Each sample from the retained UHGG sequences were merged with the de novo assembled scaffolds and the redundancy was removed at 95 % similarity using CD-HIT (v4.8.1, psi-cd-hit) [87] with parameter "-c 0.95 -G 1 -g 1 -aL 0.7 -aS 0.7 -circle 1". Prophage regions were annotated of the non-redundant representative sequences of each sample. Identification was done with "annotate" and "find-proviruses" modules in geNomad (v1.7.4) [103].

A non-redundancy representative catalog was constructed of bacteria and archaea sequences (referred to as NRbacteria from now on). For that, all sequences were combined and all redundant sequences were deleted based on a similarity threshold of 95 % utilizing CD-HIT (v4.8.1, psi-cd-hit) [87] with parameter "-c 0.95 -G 1 -g 1 -aL 0.7 -aS 0.7 -circle 1". Clean read from each sample were mapped against the NRbacteria with Bowtie2 (v2.3.5.1; configured with "--sensitive-local") [100] and SAMtools (v1.17) [101] to get the relative abundance. The relative abundance was calculated of each sample by CoverM (v0.6.1) with the specified parameters "-m tpm covered_bases length" (B. Woodcroft, unpublished, https://github.com/wwood/CoverM). NRbacteria completeness and contamination were checked using CheckM (v1.2.2) [104] 'lineage_wf' pipeline. NRbacteria were also tested for CRISPR via CRISPRidentify (v1.2.1) [105] for future viral host prediction. Taxonomical assignment of NRbacteria were done in MMseqs2 (v13.45111; DB: Swiss-Prot) [106] and Kraken2 (v2.1.2; DB: MiniKraken_DB_8GB) [107] with the default setting.

Next to the NRbacteria catalog, an NRprophage catalog was also created with all non-redundant representatives of prophage sequences. The same strategy was applied as for the NRbacteria: CD-HIT (v4.8.1, psi-cd-hit) [87] with parameter "-c 0.95 -G 1 -g 1 -aL 0.7 -aS 0.7 -circle 1". The clean reads were mapped from each sample to NRprophage with Bowtie2 (v2.3.5.1, "--sensitive-local") [100] and relative abundance were calculated using CoverM (v0.6.1; parameter: "-m tpm covered_bases length"; B. Woodcroft, unpublished, https://github.com/wwood/CoverM).

Virome

Virome raw reads were cleaned from (11 samples in total) PhiX (reference: NC_001422.1) and human (reference: GRCh38) contamination using Bowtie2 (v2.3.5.1) [100] with parameter "--sensitive-local" and SAMtools (v 1.17) [101]. In the next step, reads were cleaned from low-quality reads using fastp (v0.23.2) [78] with parameter "-z 4 -n 10 -l 60 -5 -3 -W 4 -M 20 -c -g -x". After that, we assembled clean reads individually with metaSPAdes (v3.15.2) [79] and default setting and kept everything that was longer than 1 kbp. Afterward, reads were mapped to the assembled contigs using Bowtie2 (v2.3.5.1) [100] with the parameter "--sensitive-local" and SAMtools (v 1.17).

Viral sequences were further identified with VirSorter (v1.0.6) [85] to determine putative virion contigs (VirSorter categories 1, 2, and 3) using both database options: -db 1 (RefSeq viruses) and -db 2 (RefSeq viruses+viromes). After that CheckV (v0.8.1; end-to-end) [84] was applied to delete the bacterial regions of assembled scaffolds.

In addition to the NRbacteria and NRprophage catalogs, an NRvirome catalog was also created with non-redundant viral contigs. The used viral contigs were identified via VirSorter and were processed with a dereplication procedure at a threshold of 95 % similarity. Mash (v2.3) [108] calculated the pairwise distance with a with a kmer size of 21 and a sketch size of 10,000. An in-house Mash script (mash_clstr.py) was used to determine clusters based on Mash distance. For each cluster, the longest sequence was used as a representative for that cluster. Dereplicated long viral contigs were taken to map the clean reads with Bowtie2 (v2.3.5.1) [100] with parameter "--sensitive-local" and SAMtools (v 1.17) [101].Additionally, CoverM estimated the coverage (v0.6.1; parameter: "-m tpm covered_bases length"; B. Woodcroft, unpublished, https://github.com/wwood/CoverM). Viral contigs were also taxonomically annotated with the 'annotate' function in the geNomad (v1.7.4) [103] tool and checked for the host with iPHoP (v1.1.0) [86] for host prediction.

Viral Tagging

fastp trimmed reads(v0.23.2) [78] with parameter "-z 4 -n 10 -l 60 -5 -3 -W 4 -M 20 -c -g -x". metaSPAdes assembled them (v3.15.2) [79] with the default setting, and everything longer than 1 kbp was kept for downstream analysis. Bowtie2 (v2.3.5.1, parameter "--sensitive-local") [100] and SAMtools (v 1.17) [101] were used for calculation. CheckV (v0.8.1; end-to-end) [84] and VirSorter (v1.0.6), [85] identified the viral contigs. A fourth catalog was created with viral tagging contigs (NRVvt) at a threshold of 95 % similarity using Mash (v2.3) [108]same settings as above. Clusters were determined via the in-house script (mash_clstr.py) based on Mash distance. NRVvt were mapped to clean reads using Bowtie2 (v2.3.5.1, the parameter "--sensitive-local") [100] and SAMtools (v 1.17) [101].

Cross-Assembly of Viromes and Viral Tagging

Virome samples were cross-assembled with the viral tagging samples using metaSPAdes (v3.15.2) [79] with the default setting. Next, cross-assembled contigs were dereplicated based on four disease groups at a threshold of 95% similarity using Mash (v2.3; parameter: -k 21 -s 10000) [108]. For clustering, the in-house script (mash_clstr.py) was used based on Mash-distance that also determined the representative.

Identification of Viral Sequences and Viral Clusters From Cross-Assemblies

The public human gut phage database (GPD) [109] were supplemented cross-assembled contigs. First, the non-redundant cross-assembled sequences were selected that had at least one viral gene according to CheckV (v0.8.1; end-to-end) [84] output. Second, viral sequences from NRVvirome and NRVvt were aligned separately using BLASTn (v2.13.0) [110] with parameter "-max_target_seqs 10". to the non-redundant contigs of four disease groups. Cross-assembled contigs were kept with these criteria: query coverage (qcov) was bigger than 60, and the percentage of identity (pct_identity) was more than 85. Additionally, GPD [109] was also used to support the assembly process where viral tagging reads and clean reads were mapped to with Bowtie2 (v2.3.5.1) [100] with the parameter "--sensitive-local" and SAMtools (v 1.17) [101]. GPD sequences were merged into the viral source if the read coverage of a sequence was more than 60 or the cover base (covbases) was higher than 10k.

For diversity determination, duplicates of viral sequences were eliminated. With BLASTn (v2.13.0) and the scripts from the CheckV repository (https://bitbucket.org/berkeleylab/checkv/src/master/), viral clusters were identified with (NRVcross-assemble) meeting the 98 % pairwise ANI (average nucleotide identity) and 85 % minimum coverage criteria.

Host Prediction of Cross-Assemblies

Host predictions were based on CRIRPS and host-predicting software. Host information was built on the public CRISPR database, CrisprOpenDB (download date: 202404, with default database:
http://crispr.genome.ulaval.ca/dash/PhageHostIdentifier_DBfiles.zip) [111]. BLASTn (v2.13.0) [110] with parameter "-max_target_seqs 1000" was used to align non-redundant contigs to CRISPR from NRbacteria. Sequences were selected when they had a percentage of identity (pct_identity) greater than or equal to 98 and mismatches (n_of_mismatches) equal to or less than 2 as validated matches. If CRISPR matched with the bacteria, that bacteria was used as the bacterial taxonomy. If CRISPR did not work, iPHoP (v1.1.0) [86] was used with a confidence score greater than 95. All ambiguous host genus names were manually deleted for example: "UMGS680",

Phage-Host Network Analysis

Clean reads of viral tagging samples were mapped to NRVcross-assemble and NRbacteria using Bowtie2 (v2.3.5.1, the parameter "--sensitive-local") [100] and SAMtools (v 1.17) [101] to get the viral and bacteria sources separately. The composition of all viral clusters was examined to determine which viral cluster belongs to which disease condition. The clustering was visualized using the Python package UpSet (v0.9.0)[112].

The R package NetCoMi (v1.1.0) [113] handled all network associations with its calculation and visualization.

The centered log-ratio transformation (clr) normalization method was performed with "sparcc" as the measure of associations. For simplicity, network edges were removed if the threshold was below 0.95.

Statistics

Non-parametric tests were performed for relevant statistical analyses, such as the Wilcoxon signed-rank test, with the "wilcoxon" function of stats module in the Python package SciPy (v1.14.1) [114].

Acknowledgment

In Project 1, Magdalena Unterer designed and executed the experiments, analyzed the data, and drafted the manuscript. Prof. Dr Li Deng and Dr. Mohammadali Khan Mirzaei supervised the project and supported the writing process.

For Projects 2 and 3, Magdalena Unterer performed all experiments and interpreted all results and visualizations. Dr. Mohammadali Khan Mirzaei did sample amplification and supported the bioinformatic analysis that Xue Peng executed.

2.4 Material

Key Resource Table

Reagent or Resource	Source	Identifier
Biological Samples		
Human Stool samples	This study	n/a
Patient Stool Samples	Prof. Klaus-Peter Janssen	n/a
Wastewater	Wastewater plants in Munich, Augsburg, Innsbruck	n/a
Chemicals and Media		
DNase I	thermoscientific	EN0521
Proteinase K (20 mg/mL)	invitrogen	AM2548
Syto9	ThermoFisher	S34854
Rapid Urease Test Broth	HIMEDIA	M1828-500g
Helicobacter pylori Selective Supplements	Oxoid	SR0147E
Brain-Heart-Infusion Broth	ROTH	X916.2
Wilkins Chalgren Anaerobic Broth Base	HIMEDIA	M863-500G
Horse Blood	Oxoid	SR0050C
Luria Bertani Broth	Carl Roth	X968.1
SybrGold	ThermoFisher	S11494
Agilent Brilliant Probe Multiplex MM	Agilent	600553
Water Milli-Q	Merck	n/a
Agar-Agar, Kobe I	Roth	5210.2
PCR Grade Water	VWR	733-2573
Consumables		
AMPure XP Beads	Beckman Coulter	A63881
Vivaspin 20 Ultrafiltration Unit	Sartorius	VS2042
0.22 µm Syringe Top Filter	Merck	SLGP033RS
0.45 µm Syringe Top Filter	Merck	SLHV033RS
Amicon Ultra Centrifugal Filter	Merck	UFC910024
b.sight cartridge	Cytena	
CytoFLEX Daily QC Fluorospheres	Beckman Coulter	B53230
Safe-Lock Tubes 0.5 mL	Eppendorf	0030 121.023
Safe-Lock Tubes 1.5 mL	Eppendorf	0030 123.328
Safe-Lock Tubes 2.0 mL	Eppendorf	0030120.094
CEllstar Tubes, 50 mL	Greiner bio-one	227 261
10 µL Filter Tip	Starlab	S1121-2710

Reagent or Resource	Source	Identifier
20 µL Filter Tip	Starlab	S1120-1710
200 µL Filter Tip	Starlab	S1120-8710
1,000 µL Filter Tip	Starlab	S1126-7710
Mx3000 96-Well Plates	Agilent	401334
96-Well Assay Plate	Corning COSTAR	3367
Omnifix Syringe 5 mL	Braun	4617053V
Omnifix Syringe 20 mL	Braun	4617207V
Omnifix Syringe 30 mL	Braun	4617304F
	Oligonucleotides	
Primers and Probes	Supplementary Table 1 Supplementary Table 2	
	Commercial Kits	
Repli-G Kit	Qiagen	XX150343
Genomic DNA Clean & Concentrator Kit	Zymo Research	D4064
E.Z.N.A Soil DNA Kit from Omega	Omega	D5625-02
Norgen Phage DNA Isolation Kit	Norgen Biotek Corp.	46800
DNeasy PowerLyzer Microbial Kit	Quiagen	12255-50
	Software and Data Analysis	
FlowJo v10.8.1	BD	n/a
Prism Version 8.4.3	GraphPad	n/a
BioRender	BioRender	n/a
	Machines and Equipment	
Plate Reader	BioTek EPOCH2	
MoFlo XPD	Beckman Coulter	
FACSMelody	BD	
Nanoparticle Tracking Analysis	Malvern Panalytical	
Mx3000P qPCR System	Agilent	
Thermo Mixer 13687713	Thermo Scientific	
CO2 Incubator HERACell 150i	Thermo Scientific	
b.sight	Cytena	
Vortex Genie 2	Scientific Industries	
Centrifuge 5430 R	Eppendorf	
Centrifuge 3K18	Sigma	

3. Results

This chapter describes in detail the variety of new insights that were gained throughout the development of single-cell technology in the field of phage biology and the hunt for novel phages and interactions within the gastrointestinal tract. Section 3.1 Development of Targeted Single Phage Isolation extensively describes the establishment and adaptation of single-cell technology and phages with human pathogens. The results of this research have been published in Unterer et al. [75]. The next section shows hands-on results from the previously developed technology and how it can be used to find phages for complex bacteria and extend phage reference databases in parallel. The last section took the new technology was used to analyze host-phage interactions in healthy and diseased individuals and detect key players. These new findings laid the foundation for further research, which is currently being done in Prof. Deng's lab.

3.1 Development of Targeted Single Phage Isolation

Viral tagging has been proven to be a valuable tool to link phages to their hosts and elucidate entire phage-host networks. Additionally, viral tagging also helps to isolate (new) phages. To further expand that knowledge, I established a universal viral tagging protocol that can be used for various bacterial species as well as on flow cytometers and microfluidic devices.

3.1.1 General Viral Tagging Protocol Establishment

The protocol had to work for a variety of aerobic and anaerobic bacteria in monocultures, for different phage sources and for samples containing a mixture of the unknowns. I based the first protocol attempts on the published viral tagging protocol from Džunková et al. [115], realizing that this protocol is limited as an on-site flow cytometer is necessary and that the outcome of delicate bacteria is skewed. The sample preparation time is around six hours and includes steps that put a lot of stress on the cells, resulting in cell damage and poor resolution. To minimize the stressors and time, the washing procedure needed changing. Therefore, I took the original protocol from Deng et al. [74] and merged both publications with new elements to create a more general viral tagging protocol. I created a protocol that takes less than an hour for tolerant bacteria such as *E. coli* and less than two hours for more complex organisms.

The initial development was done with *E. coli* 11303 and its phage T4 as the model organisms. Subsequent proof-of-concept runs were then performed with a variety of different bacteria, such as *Pseudomonas aeruginosa, Acinetobacter baumannii, Klebsiella pneumonia, Staphylococcus spp, Helicobacter pylori, Akkermansia municiphila* but also with stool samples from mice and human (Figure 12). To further test the robustness of the universal protocol, I tested it on multiple flow cytometer (Beckman Coulter, Sony, BD) analyzers and sorters and recreated the same results over and over again independently of the system.

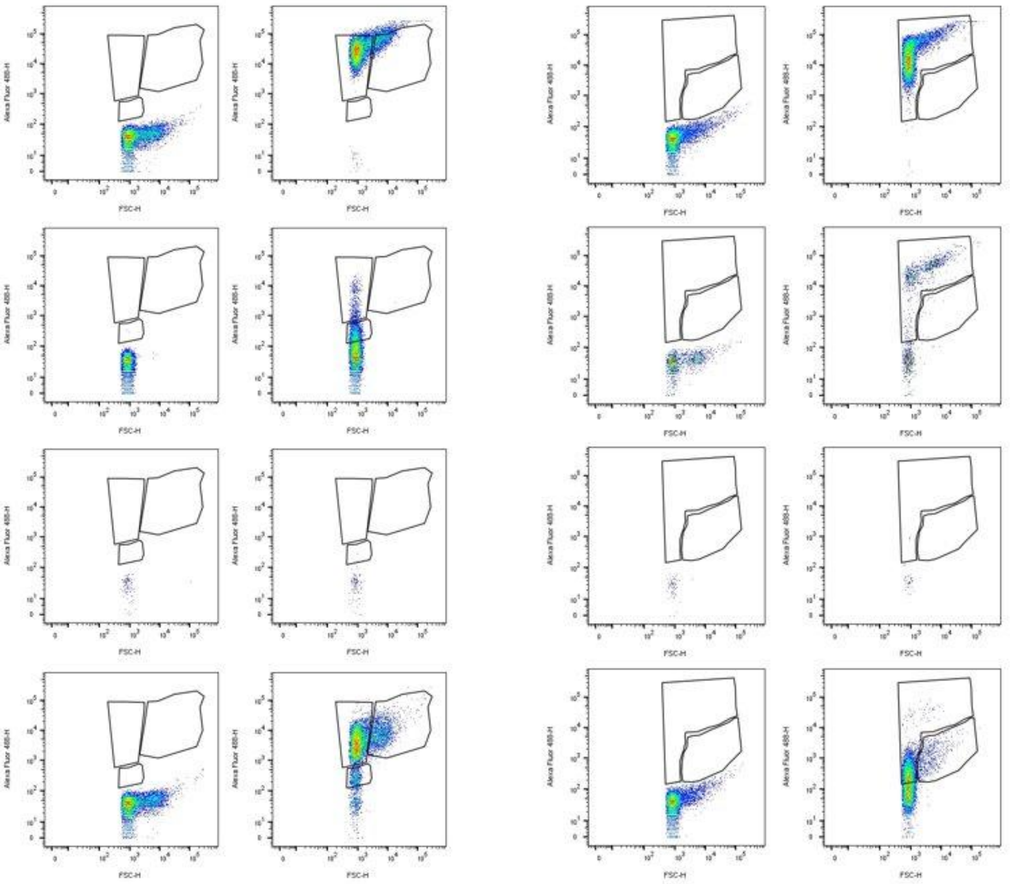


Figure 12_Viral-Tagged *Acinetobacter baumannii* (left) and *Haemophilus influenzae* (right)

Flow cytometry plots are shown. The X-axis is FSC, and the Y-axis is Alexa, the fluorescence channel. Plots are shown on an exponential scale. The first row shows unstained and stained bacterial cells, whereas unstained and stained VLPs are seen in the second row. The third row presents an unstained and stained buffer. In the last row, the left plots are unstained bacterial cells with unstained VLPs, and in the right plots, there are unstained bacterial cells with stained VLPs – the actual viral-tagged sample. Gating strategies are different between those two samples as the viral tagging protocol was in development.

3.1.2 Transition From Flow Cytometry to Microfluidics

As a next step, I transferred the entire viral tagging procedure from a flow cytometer to a microfluidics system. The reasons behind that major step are the following: flow cytometers use an external fluidics system, which can influence our phage-host system. Furthermore, the system does not allow anaerobic work. As a second point, flow cytometers work with forces; pressure is put onto cells during stream alignment and can initiate stress reactions. The chosen microfluidics system eliminates both factors. First, it works with cartridges which are a closed system without any external liquid carriers. Secondly, the microfluidics system is based on gravity, putting little to no pressure on the sample. Compared to the flow cytometers, the microfluidic system identifies cells not via light scattering but via cameras. The system detects the cells in a brightfield and fluorescence mode. In Figure 13, a stained cell was detected, first with the brightfield camera, then with the fluorescence camera, and the third picture is the overlay of both. The cameras capture the cell and determine the size and the fluorescence value. Depending on the sort settings, cells are then dispensed or not.

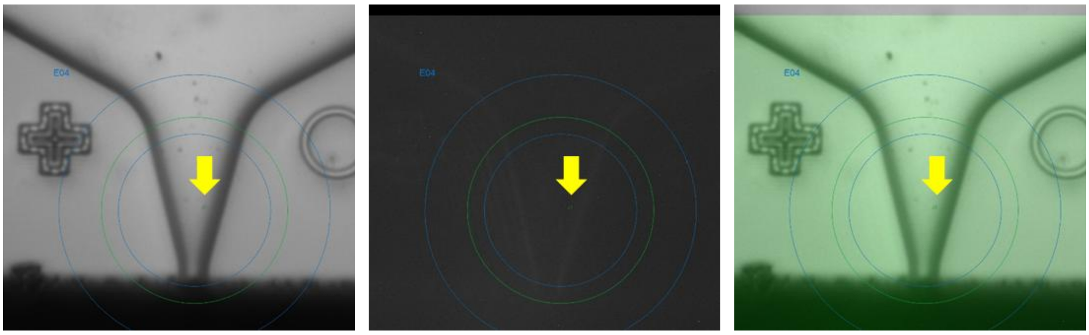


Figure 13_b.sight Camera Channels

All three photos show the same cell marked with the yellow arrow but in different channels. The far left picture is the brightfield mode, the middle one is the fluorescence channel, and the far right photo is an overlay of both. The circles inside are defined areas by the software to ensure only a single cell is sorted.

Since the microfluidics device was customized to our needs, beta testing was necessary to achieve the transition. The appropriate amount of events per second on flow cytometers is set and known and can, therefore, be targeted when samples are prepared. For the microfluidics, I had to test that. To gain first insights into how our bacteria appear on that camera and at which concentration the cartridge can be challenged, I started with unstained and stained *E. coli*. I found that a bacterial concentration of about 10^5 cfu/mL is suitable to keep the cartridge unblocked and maintain a smooth flow. A higher bacterial concentration placed the cartridge at risk for blocking, and sorting was delayed because too many cells were detected within the sort area, which made single-cell sorting impossible. A bacterial concentration that was too low delayed the process as well since too few cells were captured, and the proximity to finding the correct cells was low, resulting in a slow sort rate. Four consecutive pictures (Figure 14) are shown where each cell is marked with an arrow while it flows through the cartridge. Since cells are flowing unevenly within the cartridge, the right amount is important to find the balance between the risk of blocking, too few cells and being fast.

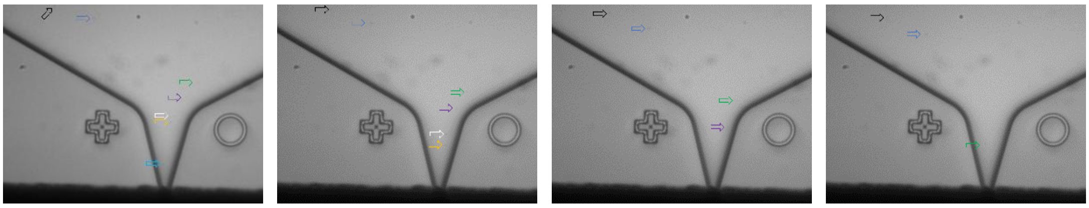
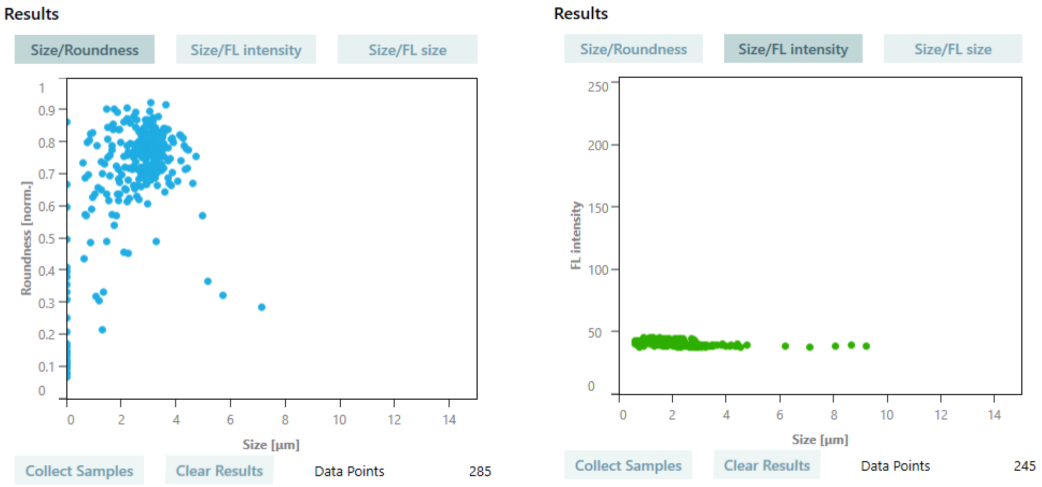


Figure 14_Cell Movement Inside b.sight Cartridge

Cells are marked with different colored arrows while moving through the b.sight cartridge. It is a series of five pictures.

After determining the right bacterial concentration, I studied the graphical output for bacterial cells regarding size and roundness, which are the comparable parameters of FSC and SSC on flow cytometers. Bacterial cells have been found to show up between 0.8 and 4 in size depending on bacterial species and between 0.5 and one for roundness. As a next parameter, fluorescence was determined. Autofluorescence was recorded for unstained bacterial samples and minimum fluorescence for stained bacterial cells. Graphical outputs were studied from different species. I found that autofluorescence has its maximum at 45 and below throughout different species regardless if they are aerobe or anaerobe, and stained bacterial cells had their minimum around 55 and above and were highly distinguishable from non-stained cells (Figure 15.).



Data Points 285

Data Points 245

Figure 15_Graphical Output of b.sight

The left picture represents the cell's shape, whereas the right picture displays their fluorescence level. Axes are size, roundness and fluorescence on a linear scale.

The generalized previously established viral tagging protocol was tested after these parameters were established and used as a reference point. Viral-tagged *E. coli* with T4 were placed inside the cartridge and monitored. Bacterial cells were nicely distributed with the chosen concentration however, I could not sort properly as fluorescence was constantly detected within the sorting area. As a result of that, the phage concentration was adjusted, and a sample ratio of 1:2 (bacteria: phage volume) was found to be sufficient. After the adjustment viral-tagged *E. coli* was tested again and compared to flow cytometer results. The results can be seen in Figure 16. Cells scattered in the same pattern independently of the machine, and viral-tagged cells were distinguishable from non-tagged cells depending on their fluorescence. Independently, if viral-tagged cells are defined via gating on flow cytometers and via parameter settings on the b.sight.

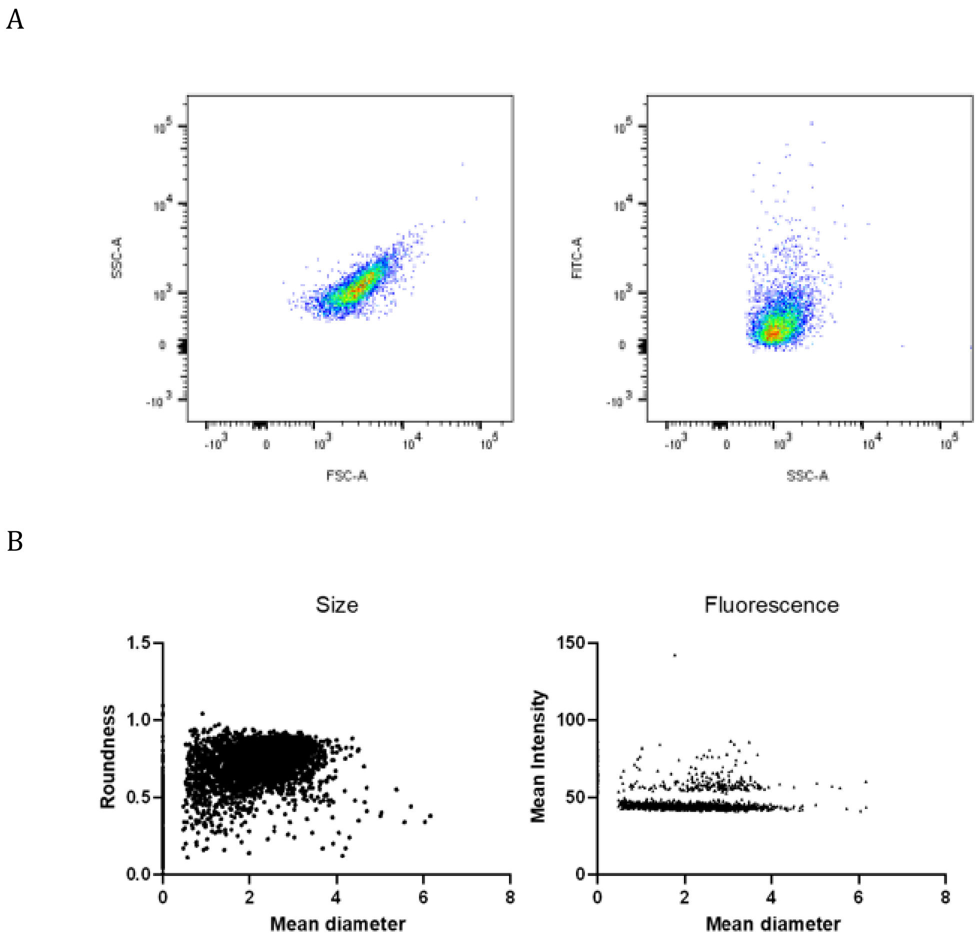


Figure 16_Viral-Tagged Sample Shown on Flow Cytometer vs. on b.sight (microfluidics)

All four plots are of viral-tagged cells, but the upper two (A) are from the flow cytometer, and the lower two (B) are from b.sight. The left graphs show cell size and shape, and the right plots cell fluorescence as presented by axes.

After visual confirmation that the established protocol works, viral-tagged cells were sorted. Since no gating strategy can be used on b.sight, the range between autofluorescence and minimum fluorescence was used as sorting parameters. Viral-tagged cells were sorted onto a layer of soft agar and bacterial suspension ranging from 1 to 100 cells per position/well. Single-tagged cells could not be seen on agar but the highest amount of sorted cells did produce visible plaques, confirming the presence of phage. In contrast, each well of the bacterial suspension had some sort of phage activity present, confirming the presence of phages there and the successful transition of the viral tagging protocol from flow cytometer to microfluidics.

3.1.3 Phage Titer Determination

In the next phase, I had to ensure enough DNA was available for a sufficient PCR run. Therefore, I tested different bacterial concentrations to see how the phage titer differs and if the phage titer provides enough DNA for the PCR. I serially diluted a known concentration to get *E. coli* 11303 at 10^6 cfu/mL, 10^5 cfu/mL, 10^4 cfu/mL, and 10^3 cfu/mL. One single viral-tagged phage-host pair was sorted into each well. Negative controls were included for each condition: bacteria only and media only.

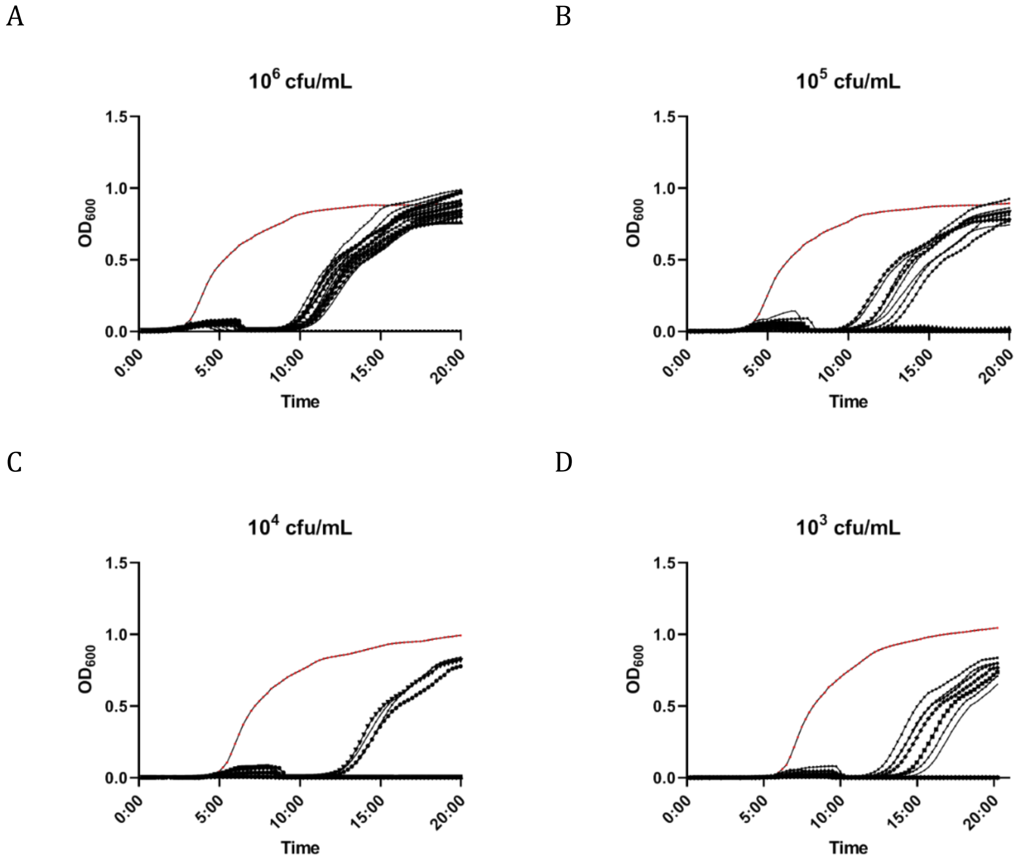


Figure 17_Single-Cell Sorting Into Different Bacterial Concentrations

One single viral-tagged cell was sorted into wells containing different bacterial concentrations. The red curve represents bacterial growth without viral-tagged cells to the corresponding bacterial concentration. Black curves are samples including one viral-tagged cell. Plots are presenting the time on the x-axis and OD_{600} on the y-axis. The upper left corner (A) shows samples with 10^6 cfu/mL and the trend that with that bacterial concentration, bacterial regrowth happens. Less of this regrowth is seen in the upper right corner (B), which displays growth graphs from samples with 10^5 cfu/mL. The lower left graph (C) includes curves from samples with 10^4 cfu/mL displaying the least amount of bacterial recovery. Results with 1,000 cells and phage are seen in the lower right graph (D), which shows more bacterial regrowth again.

Results are shown in Figure 17. All bacterial-only controls show growth curves as expected, differ only at their start time for the exponential phase based on their concentrations; the higher the concentration, the earlier the exponential phase started, and the lower the concentration, the later the growth. Medium-only control was negative throughout all experiments. As for the samples themselves, they show growth and killing activity. 21 samples with 10^6 cfu/mL had two wells where bacteria were killed off completely and 19 wells with bacterial regrowth after phage-killing. The result was half-half for samples with one log less bacterial cells (21 wells). Ten wells showed bacterial regrowth, and 11 no longer had bacterial activity. Again, a different picture was seen for 10^4 cfu/mL samples. Out of 21 wells, only five showed signs of bacterial regrowth, whereas 16 wells had a total bacterial loss. Against the trend, out of 15 wells with 1,000 cells, seven had bacterial regrowth, and eight had none. To further test if phages were present and, if yes, how many, the supernatant was harvested. 10 µL of serially diluted samples were spotted onto a soft agar layer and incubated overnight. All four conditions produced full lysis for the first five dilutions, and after that, single plaques were formed (Figure 18). For the bacterial concentration, 10^6 cfu/mL and 10^5 cfu/mL, a phage titer of 10^9 pfu/mL was calculated, and for the two lower concentrations, one log less. However, both titers are enough to extract DNA and perform PCRs.

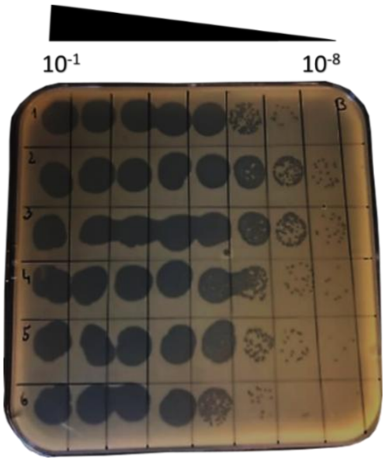


Figure 18_Representative Spot Assay to Determine Phage Titer

Eight different concentrations were tested, with the highest concentration on the left and the highest dilution on the right. The first five formed spots with total lysis, and further dilutions produced single plaques.

The aspiration was to create a highly standardized, reproducible, and less operator-biased protocol. I tested further bacterial concentrations, but instead of diluting serially, which is highly prone to operator failure, I sorted a certain amount of bacterial cells using the microfluidics device. Ten, five, and one cell(s) per well were tested, and bacterial growth in all wells was achieved only with tens cells. In the next run, viral-tagged cells were added and monitored again. Figure 19 shows in red the bacterial-only curve and in black all different wells with phage activity.

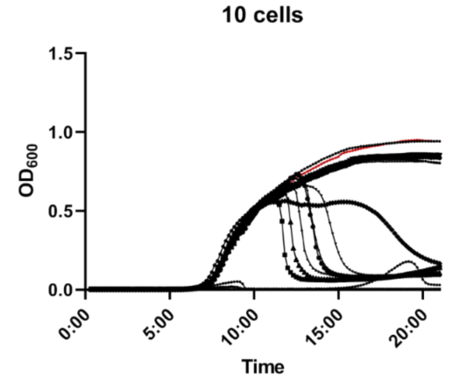


Figure 19_10 Cells and 1 VT Cell

Ten bacterial cells were sorted together with one viral-tagged cell and monitored. In red, bacteria-only control is shown, and in black, test samples. The x-axis is time, and the y-axis is OD_{600}.

Plaque assays confirmed a sufficient number of phages within those samples. As a consequence, further experiments were performed with ten sorted bacterial cells as a basis.

3.1.4 Development of Multiplex qPCR

3.1.4.1 Primer Design

Despite being available in vast quantities, not all phages are suitable for phage therapy. Some of them can integrate themselves into their hosts; others transfer toxins or virulence factors. Since these features can only be found after the isolation process, all phages must be screened via sequencing or PCR. Since PCR requires known sequences, I have taken conserved regions within the phages that are known to be efficient, fast, or have a broader host range. For that, I took conserved regions from the major head protein, the conserved region for the virulence factor eae (intimin), and the *E. coli* toxin heat-labile enterotoxin. All taken regions are seen in Figure 20

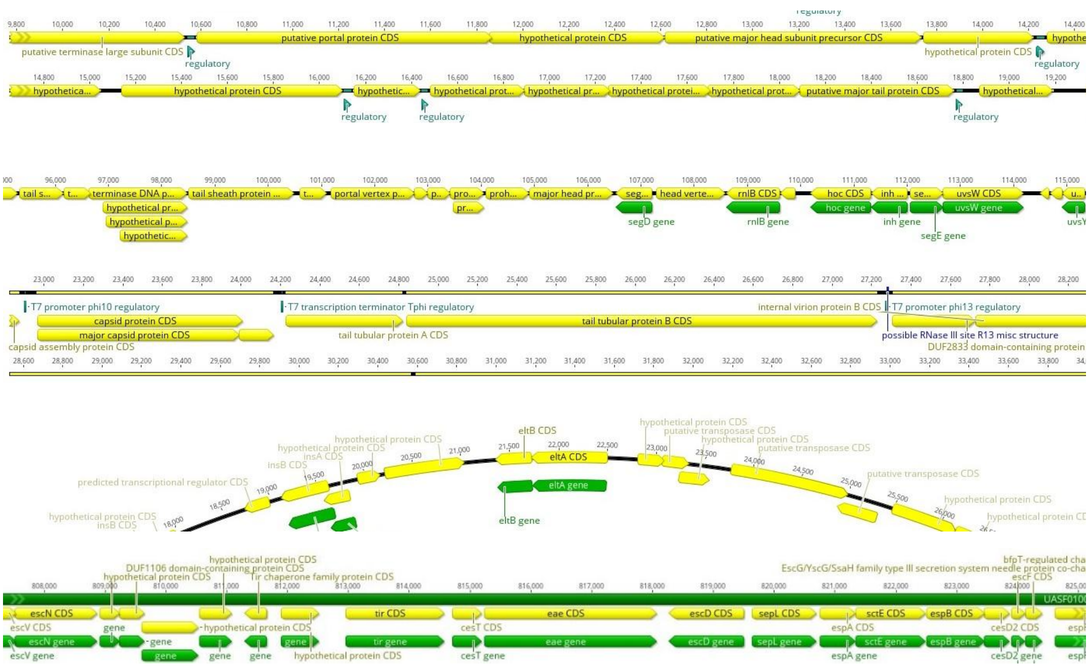

putative terminase large subunit CDS
regulatory
putative portal protein CDS
hypothetical protein CDS
putative major head subunit precursor CDS
hypothetical protein CDS
putative major tail protein CDS
tail sheath protein
terminase DNA p...
portal vertex p...
major head pr...
head verte...
rnlB CDS
hoc CDS
uvsW CDS
segD gene
rnlB gene
hoc gene
inh gene
segE gene
uvsW gene
T7 promoter phi10 regulatory
capsid protein CDS
major capsid protein CDS
capsid assembly protein CDS
T7 transcription terminator Tphi regulatory
tail tubular protein A CDS
tail tubular protein B CDS
internal virion protein B CDS
T7 promoter phi13 regulatory
possible RNase III site R13 misc structure
DUF2833 domain-containing protein
eltB CDS
eltA CDS
eltA gene
eltB gene
insA CDS
insB CDS
predicted transcriptional regulator CDS
putative transposase CDS
hypothetical protein CDS
DUF1106 domain-containing protein CDS
Tir chaperone family protein CDS
EscG/YscG/SsaH family type III secretion system needle protein co-cha
bfpT-regulated cha
escF CDS
escN CDS
escV CDS
tir CDS
eae CDS
escD CDS
sepL CDS
sctE CDS
espB CDS
espA CDS
cesT CDS
cesD2 CDS
escN gene
escV gene
tir gene
eae gene
escD gene
sepL gene
sctE gene
espB gene
espA gene
cesT gene
cesD2 gene
UASF010

Figure 20_Sequences for Primer Design

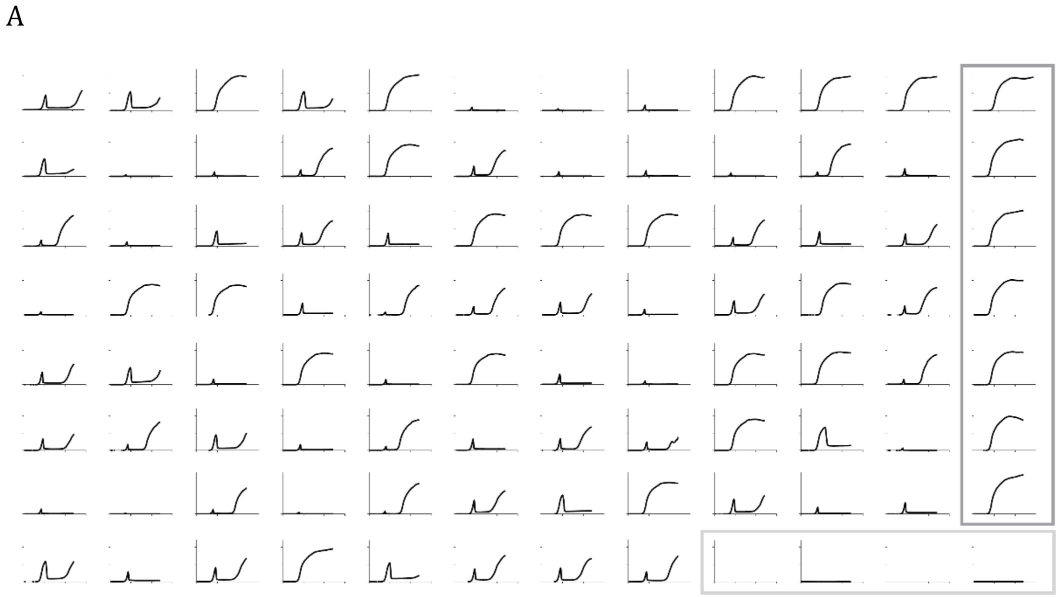
Major head genes, toxin genes, and virulence genes are shown in these graphics. Those genes were used to design primers to detect those traits in isolated phages.

3.1.4.2 qPCR Performance

After the primer design, the multiplex qPCR needed to be established. For that, DNA from phages and bacteria was extracted to have positive controls. Different primer and probe concentrations were tested singularly and were found to be at their best for 100 nM probe and 300 nM for each primer. This master mix could amplify targets within a range between one picogram to ten nanograms. To further simplify the protocol, probes, and primers were designed to work in a multiplex PCR setting. So, in the next step, I created a master mix containing all three probes, all six primers, and the reference dye and tested it. There was no amplification for wells without DNA, which confirms that none of the primers are amplifying each other and that there is no unspecific amplification. Samples with only one target DNA had one amplification specific to its target, and multiple amplifications happened in wells, including various DNAs. Any tested conditions did not influence the intensity of the signal. Since DNA extraction from each sorted phage sample is very time-consuming and costly, I tested the protocol with pure supernatant. Despite its promising looks at the start of the PRC reaction, the curve starts to decline and decreases into negative fluorescence values independently of phage type. In close collaboration with Agilent, a simple solution was found. A 1:10 dilution from the supernatant dilutes the contaminant in a way that does not interfere with the amplification process anymore. 1:10 dilutions or pure DNA perform equally well during the PCR process.

3.1.5 Proof-of-Concept Runs: Targeted Single Phage Isolation

Proof-of-concept runs were performed from start to end with *E. coli* and T1, T4, and T7. For each combination, infection dynamics were monitored. Bacterial-only growth is marked with a dark grey square, whereas media-only is highlighted in bright grey. For all three setups, bacteria grew as expected, and no contaminations happened. PCRs had no amplification for water and no template controls and amplification curves for all six different target concentrations. For each tested well, target amplification started after 15 amplification cycles when phage was present or never in the case of phage absence. (Figure 21, Supplementary Figure 1, Supplementary Figure 2)



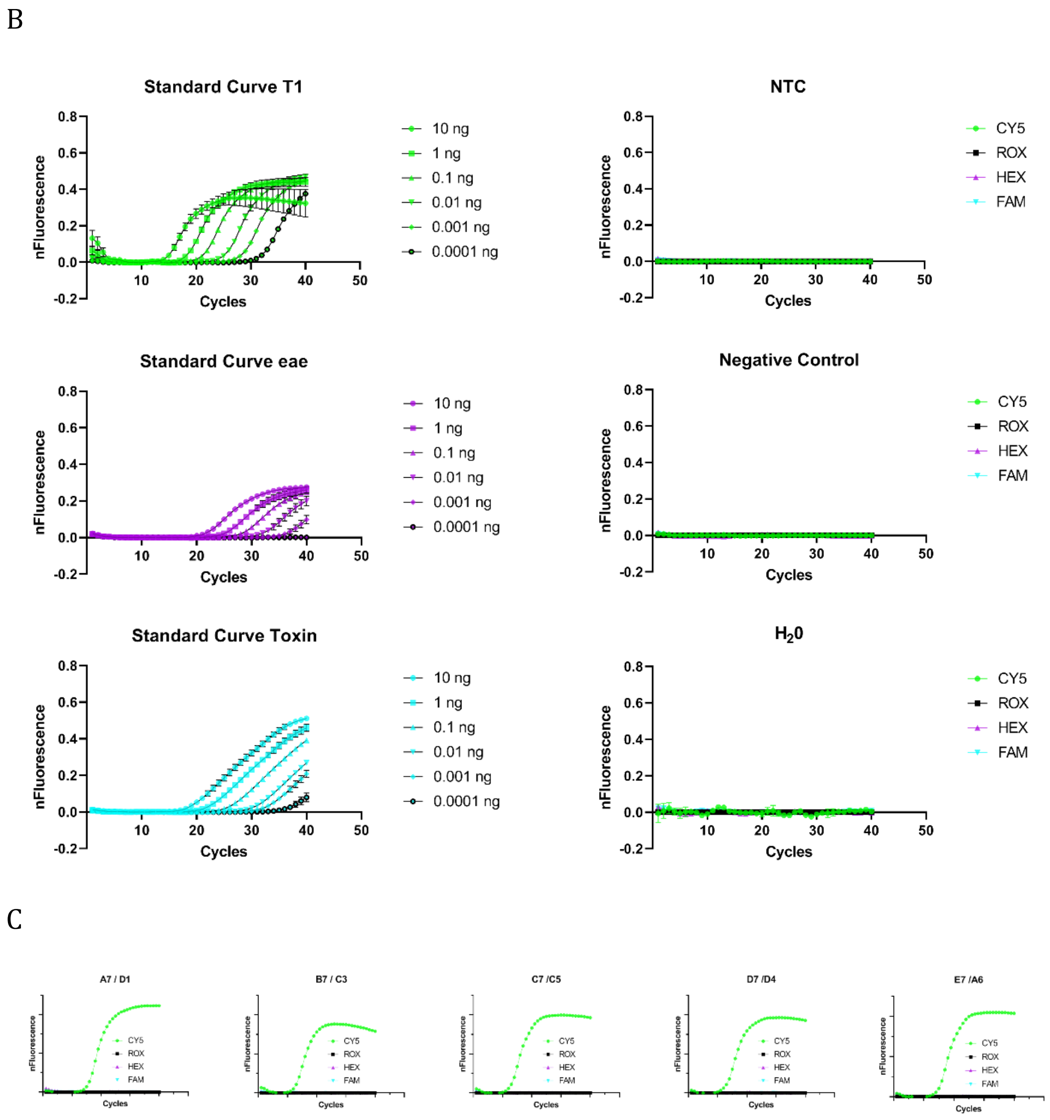


Figure 21_Proof-of-Concept Results From T1

A_Infection Dynamics were monitored overnight. Bacterial-only controls are marked with a dark grey square, whereas media-only controls are with light grey.

B_On the left, qPCR controls are shown from T1, the toxin and the virulence factors with the number of cycles on the x-axis and the normalized fluorescence on the y-axis. Six concentrations were tested from 10 ng to 1 pg. On the right, three negative controls are shown. On the top, there is the no-template control (NTC); in the middle, the negative control; and on the bottom, water only.

C_Five representatives of qPCR amplification of target samples. There are no amplifications from toxins or, virulence factors or cross-contaminations. All samples amplify around 15 cycles.

3.1.6 Sensitivity Test

To further challenge the protocol, I performed sensitivity and quality checks. First, I tested the sensitivity of the experimental process to see to what extent we can detect phages. For that the protocol was run with three different spiked conditions. Wastewater was spiked with either equal amounts of VLPs/mL, 1000x less and 1000x more. All three model phages were used for spiking. I always evaluated infection curves from the plate reader, qPCR data, and spot assay results. When I had more wastewater than phages, I detected no phages in any of the three methods. The same was true for T1 and T7 for equal amounts. T4 had two positive wells when there were equal amounts of phages and wastewater VLPs. However, when I added 1000x more phage VLPs than wastewater VLPs, I had 80 positive wells for T1, 10 for T4, and two for T7, whereas 15 showed active infection but no signs during qPCR amplification or spot assay. Test results are summarized in Table 2.

Table 2_Spiking Data

	Sample	Plate Reader Kinetics	qPCR	Spot Assay
T1	More WW VLPs	0 wells	0 wells	0 spots
	More T1	80 wells	80 wells	80 spots
	Equal amounts	0 wells	0 wells	0 spots
T4	More WW VLPs	0 wells	0 wells	0 spots
	More T4	10 wells	10 wells	10 spots
	Equal amounts	2 wells	2 wells	2 spots
T7	More WW VLPs	0 wells	0 wells	0 spots
	More T7	15 wells	2 wells	2 spots
	Equal amounts	0 wells	0 wells	0 spots

To check if these results are a lack of sensitivity from the test procedure or if this is the result of single-cell sorting, I sorted one million cells with the same conditions as prior using T1 as a representative. Flow cytometry plots were compared, and the more T1 VLPs were included in the sample, the more viral-tagged cells appeared inside the target gate. For the sample with the least T1, the percentage inside the gate was 2.59 %. For the sample with equal amounts of VLPs, the gate held 26.2 % of the cells, and for the sample with the most T1, 38.5 % of cells showed up inside the gate. With one million sorted cells, T1 could have been recovered from those three conditions during qPCR runs. The amplification curve of T1 DNA-only is shown in green and starts around 15 cycles, as it was established during the multiplex qPCR development. The sample with more T1 VLPs started to be amplified around 22 cycles, shortly followed by the sample with equal amounts. At 30 cycles, the sample with the least amount of T1 started to be amplified. Controls for toxin and virulence factors, as well as no template control, did not show any sign of amplification. Flow cytometry and qPCR results are presented in Figure 22.

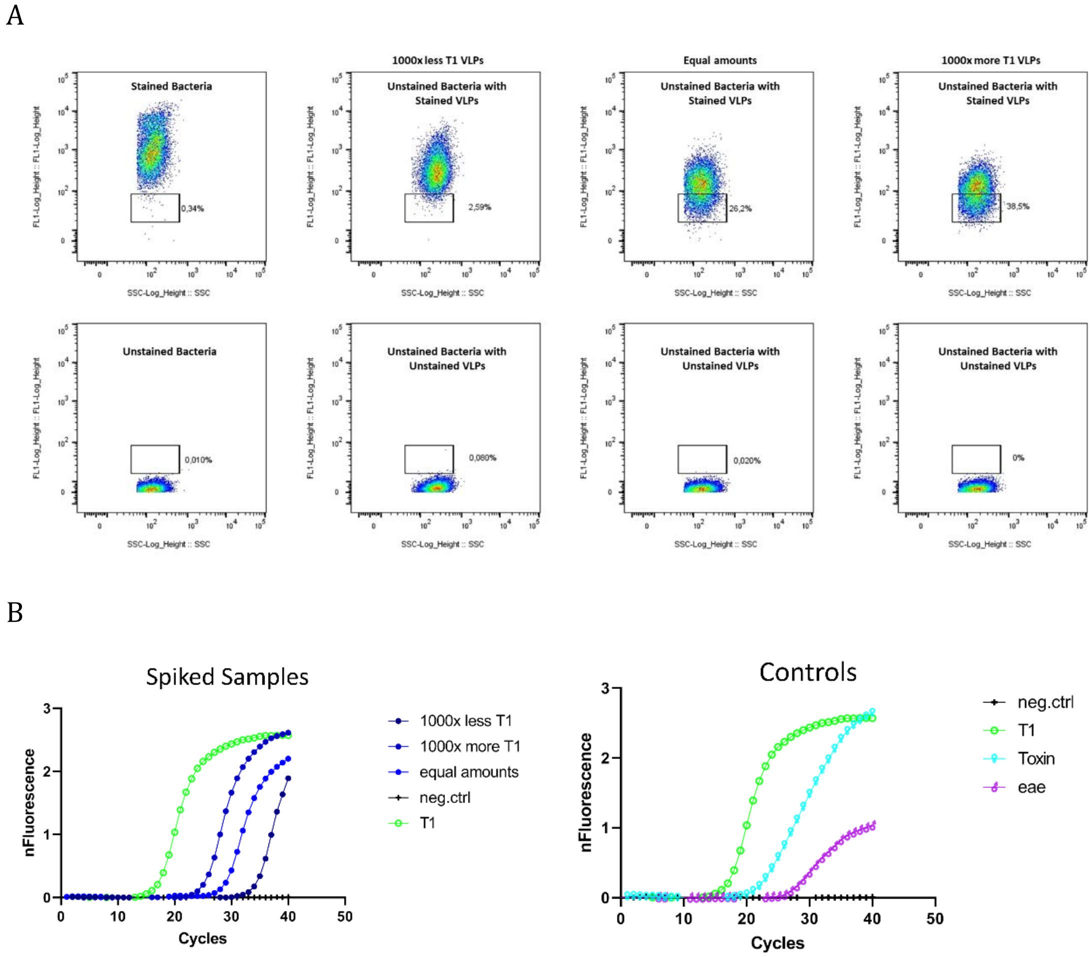


Figure 22_Spiking Results

A_Flow Cytometer plots of spiked samples. The x-axes are side scatter, and the y-axes are fluorescence on a bi-exponential scale. The top row is stained bacteria and spiked mixtures with unstained bacteria plus stained VLPs. The bottom row is unstained bacteria and spiked unstained mixtures. T1 concentration increases from left to right, increasing the amount of cells in the gate.

B_qPCR results from spiked samples where T1 was detected in all three conditions. No toxins or virulence factors were amplified within the target samples, but controls did work.

In the next step, I tested if the experimental procedure was sensitive enough to detect spiked samples and check the procedure's robustness to distinguish between different phages. Therefore, I sorted a viral-tagged sample with all three model phages included in the same amount, once on a single-cell level and once in bulk (one million cells). Only

T1 was recovered during the qPCR runs on a single-cell level, whereas when one million cells were sorted, all three phages were detected during the qPCR. The qPCR could differentiate between T1, T4, and T7 despite being mixed. Pure phage DNA as a positive control amplified around 15 cycles, and the sorted cells with less concentration roughly 10 cycles later. During flow cytometry, phages were not distinguishable since only one cloud of viral-tagged cells appeared. The results are shown in Figure 23.

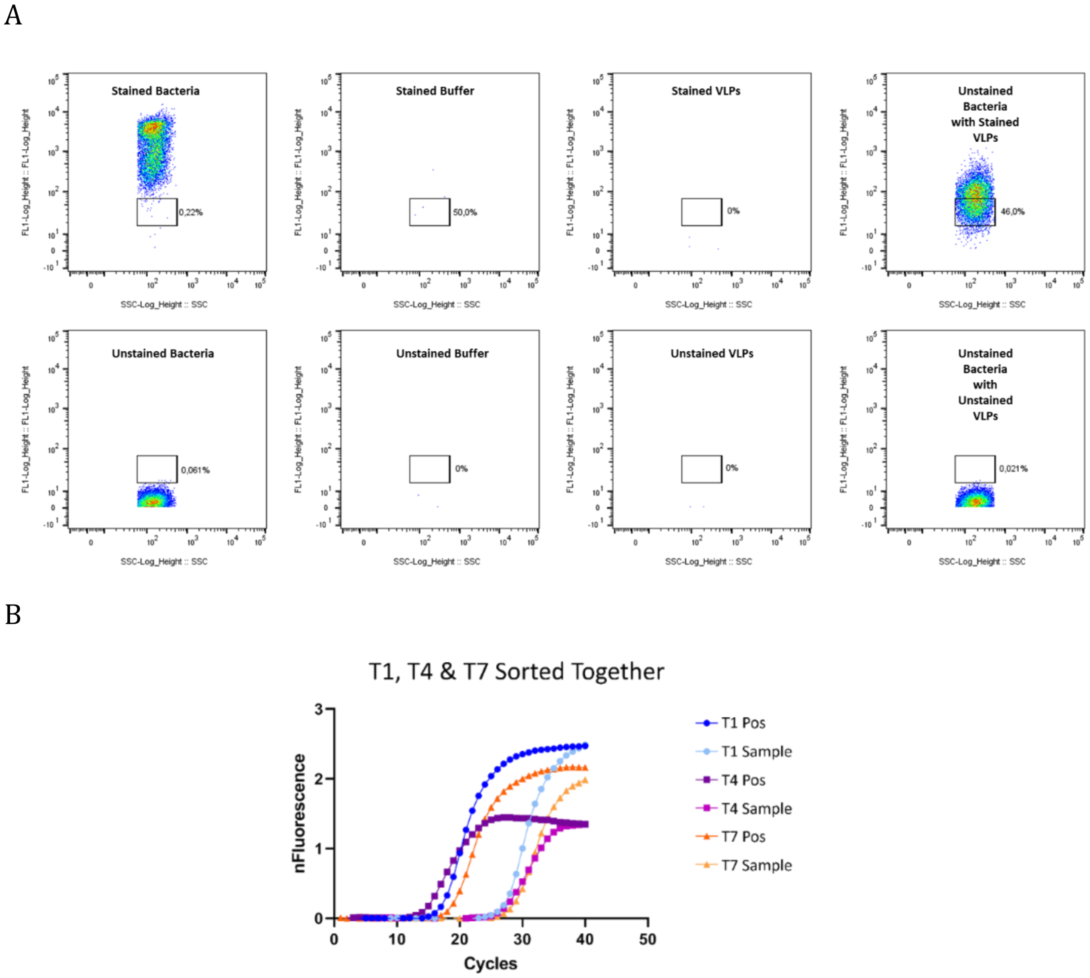


Figure 23_Robustness Test

Flow Cytometry Plots (A) show unstained bacteria, unstained buffer, unstained VLPs and unstained mixture in the bottom row from left to right. In the top row, there are stained bacteria, stained buffer, stained VLPs, and the viral-tagged sample (unstained bacteria with stained VLPs). VLPs are a mixture of T1, T4 and T7 in equal amounts. The x-axis is side scatter, and the y-axis is fluorescence. Both axes are on a bi-exponential scale.

qPCR result (B) shows that all three phages were detected, although they were in one sample. For each phage, a positive, DNA only, control was added. T1 (dots) in dark and bright blue, T4 (square) in dark and bright purple, and T7 (triangle) in dark and bright orange. Positive samples (Pos) started to amplify around cycle 15, whereas sorted cells (Sample) started ten cycles later.

3.1.7 Heterogeneous Infection Dynamics

After acquiring all those results, I took a closer look at how each phage performed in the presence of their hosts on a single-cell level, knowing that they behave differently at the spiking experiments. I analyzed the infection property and calculated the area under the curve (AUC) and found despite being the same phage with the same hosts in the same experiment, there is a cell-to-cell variety not only across different phages but also within one phage itself. I determined four distinct patterns: no infection at all, total lysis from the beginning with no regrowth of the bacteria, and two hybrid forms. The first hybrid form is that the bacteria grew until the phage infected and killed them and never regrew. The second hybrid form is similar to the first one, but the bacteria recovered and regrew (Figure 24).

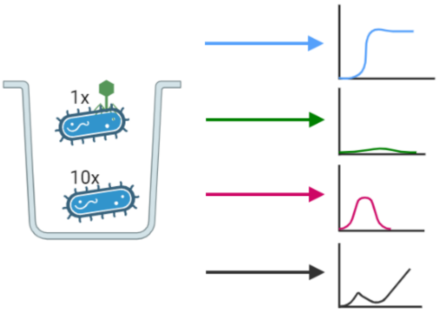
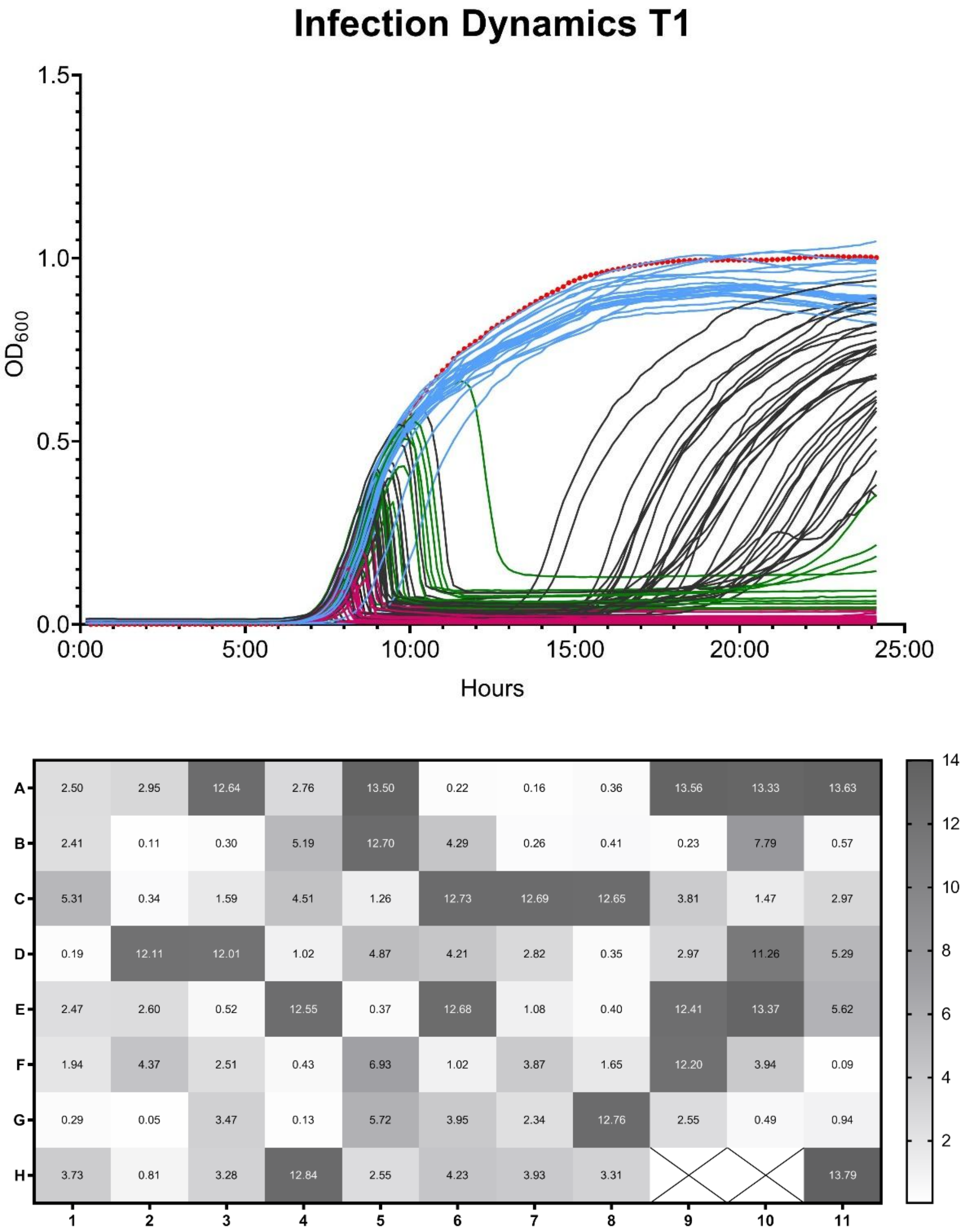


Figure 24_Phage Infection Pattern

Four major patterns were detected for single phage-host sorting. All curves in blue had no phage infection and bacteria grew. Total bacterial loss was seen in every green curve. Purple represents all samples with bacterial growth at the beginning and then phage infection. All black curves had phage activity initially, but bacteria regrew in the end.

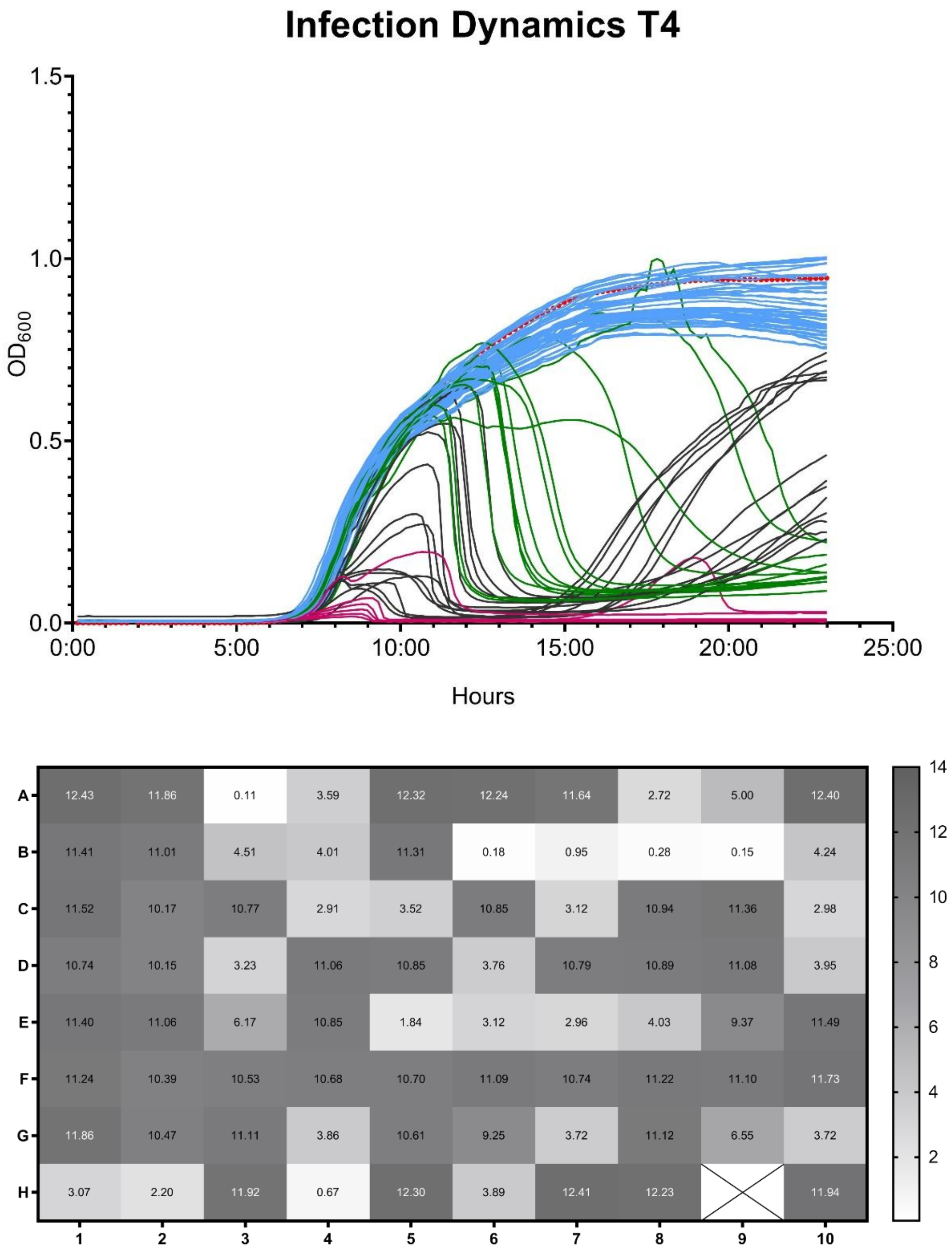
As plaque assays confirmed the presence of phages for the later three patterns, I calculated an active infection rate of 78.8 % for T1, whereas most of the wells had a low or median AUC. For T4, 42.3 % of active infections were calculated, but only six had really low areas under the curve; the remaining were medium or high. 5.9 % active infection happened for T7, leading to 5 wells with low AUCs. Data are presented in Figure 25. Backing those results with qPCR, samples with active infection were amplified, whereas no infection samples had no amplification.

A



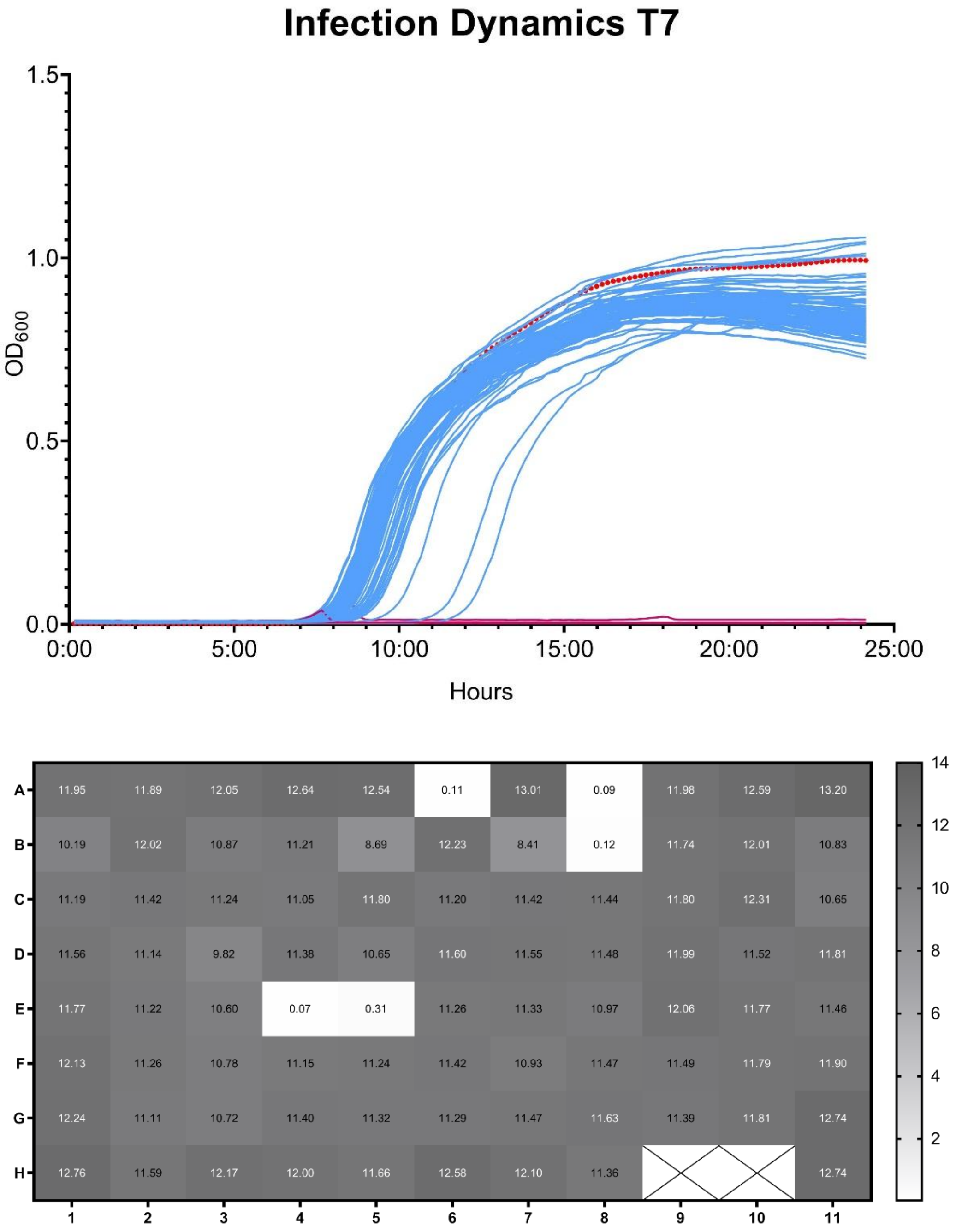
	1	2	3	4	5	6	7	8	9	10	11
A	2.50	2.95	12.64	2.76	13.50	0.22	0.16	0.36	13.56	13.33	13.63
B	2.41	0.11	0.30	5.19	12.70	4.29	0.26	0.41	0.23	7.79	0.57
C	5.31	0.34	1.59	4.51	1.26	12.73	12.69	12.65	3.81	1.47	2.97
D	0.19	12.11	12.01	1.02	4.87	4.21	2.82	0.35	2.97	11.26	5.29
E	2.47	2.60	0.52	12.55	0.37	12.68	1.08	0.40	12.41	13.37	5.62
F	1.94	4.37	2.51	0.43	6.93	1.02	3.87	1.65	12.20	3.94	0.09
G	0.29	0.05	3.47	0.13	5.72	3.95	2.34	12.76	2.55	0.49	0.94
H	3.73	0.81	3.28	12.84	2.55	4.23	3.93	3.31	☒	☒	13.79

B



	1	2	3	4	5	6	7	8	9	10
A	12.43	11.86	0.11	3.59	12.32	12.24	11.64	2.72	5.00	12.40
B	11.41	11.01	4.51	4.01	11.31	0.18	0.95	0.28	0.15	4.24
C	11.52	10.17	10.77	2.91	3.52	10.85	3.12	10.94	11.36	2.98
D	10.74	10.15	3.23	11.06	10.85	3.76	10.79	10.89	11.08	3.95
E	11.40	11.06	6.17	10.85	1.84	3.12	2.96	4.03	9.37	11.49
F	11.24	10.39	10.53	10.68	10.70	11.09	10.74	11.22	11.10	11.73
G	11.86	10.47	11.11	3.86	10.61	9.25	3.72	11.12	6.55	3.72
H	3.07	2.20	11.92	0.67	12.30	3.89	12.41	12.23	☒	11.94

C



	1	2	3	4	5	6	7	8	9	10	11
A	11.95	11.89	12.05	12.64	12.54	0.11	13.01	0.09	11.98	12.59	13.20
B	10.19	12.02	10.87	11.21	8.69	12.23	8.41	0.12	11.74	12.01	10.83
C	11.19	11.42	11.24	11.05	11.80	11.20	11.42	11.44	11.80	12.31	10.65
D	11.56	11.14	9.82	11.38	10.65	11.60	11.55	11.48	11.99	11.52	11.81
E	11.77	11.22	10.60	0.07	0.31	11.26	11.33	10.97	12.06	11.77	11.46
F	12.13	11.26	10.78	11.15	11.24	11.42	10.93	11.47	11.49	11.79	11.90
G	12.24	11.11	10.72	11.40	11.32	11.29	11.47	11.63	11.39	11.81	12.74
H	12.76	11.59	12.17	12.00	11.66	12.58	12.10	11.36	☒	☒	12.74

Figure 25_Heterogeneous Infection Dynamics

Different infection dynamics are presented as growth curves in the upper row and as area under the curve below. Red curves are bacteria only controls or shown as X in the heatmap. No phage activity are displayed in blue or dark grey. Full bacterial killings are presented in green or white. Purple, black and shades of grey are variations of infection. T1 (A), T4 (B) and T7 (C) have different infection patterns within themselves but also compared to each other.

3.1.8 Cytotoxicity Test and One-Step-Growth Curves

After gaining those insights, I wanted to rule out any influences on the cells during the experimental procedure. Therefore, I performed cytotoxicity tests and one-step growth curves to study phage behavior. Phage infectivity was not influenced by Syto9 as the titer did not decrease for any phage. So, there was no cytotoxic influence from the dye. One-step-growth curves (Figure 26) confirmed that neither the sample preparation itself nor the sorting or continuous shaking during the incubation has an effect on the cells or phage-host interaction. Phage titer decreased within the first 10 to 20 minutes until phage infection occurred. After that, new progenies were released, which led to a higher titer afterward. Absorption rate and burst size were as expected. No abnormalities were recorded for T1, T4, or T7.

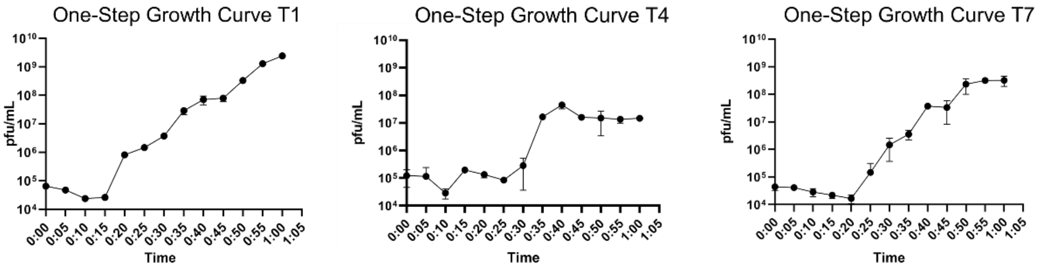


Figure 26_One-Step-Growth Curves

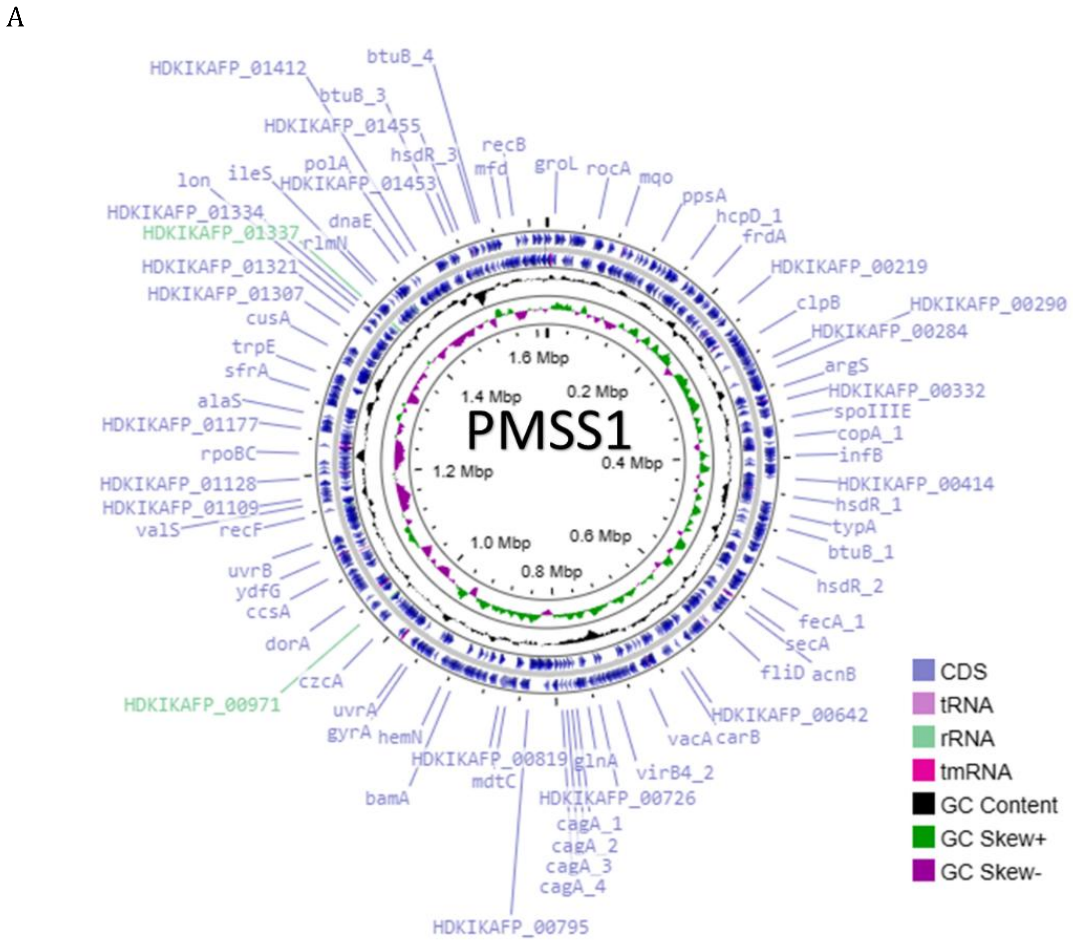
One-Step-Growth curves are shown for T1, T4 and T7. The x-axis is time in minutes and the y-axis pfu/mL. The adsorption rate is between 10 and 20 minutes before progenies are released and phage titer increases.

3.2 *Helicobacter pylori* Phages and Their Diversity

3.2.1 Genome Comparison of *Helicobacter pylori* PMSS1 and SS1

As bacteria evolve quickly in natural or lab surroundings, *Helicobacter pylori* strains were sequenced to determine whether DNA changes had happened or not.

Whole genome sequencing was performed, and the analysis revealed the correctness of both strains visualized in Figure 27 (A and B). Both have a genome length of 1.6 Mbp and a GC- Content of 39 %. Geneious whole genome alignment showed a pairwise identity of 98.1 % for the first Local Colinear Blocks (LCB), 99.6 % for LCB2 and LCB3.



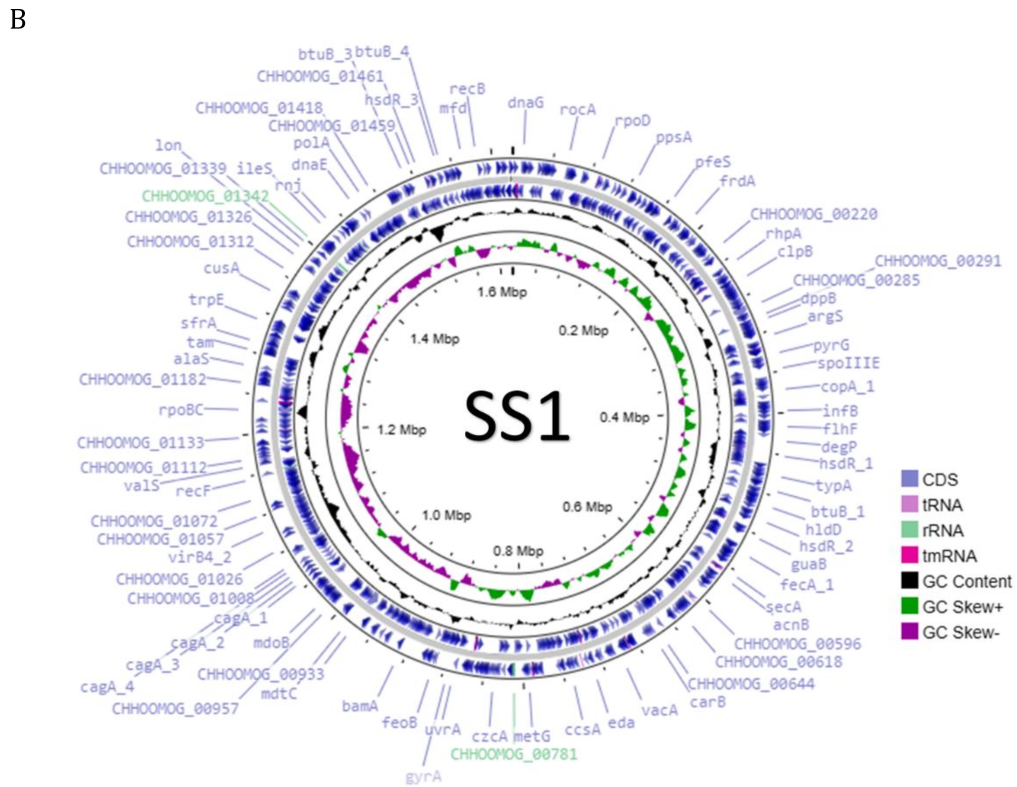


Figure 27_Genomic Maps of *Helicobacter pylori* PMSS1 and SS1

Map A shows *Helicobacter pylori* PMSS1 used during the experiments with its genetic features in blue and its GC content in black. Same representation was chosen for the genomic map of *Helicobacter pylori* SS1 (map B).

3.2.2 Phage Detection for *Helicobacter pylori*

Viral-tagged samples were sequenced and after reads were cleaned, every contig over 1,000 basepairs were kept for analysis. As a total, we found 22,331 contigs. 11,429 contigs were isolated on both strains with wastewater. 8,417 contigs belong to *Helicobacter pylori* SS1 and only 3,012 to *H. pylori* PMSS1. A different picture was seen with stool samples. 4,958 contigs were assembled from *Helicobacter pylori* PMSS1 and 4,522 from *Helicobacter pylori* SS1, slightly less than with PMSS1. However, overall, 61.88 % of the contigs belong to *H. pylori* SS1 and 38.12 % to *H. pylori* PMSS1. The exact numbers are listed in Table 3.

Table 3_Amount of Viral Contigs Found in Wastewater and Stool Samples for Helicobacter pylori PMSS1 & SS1

	Helicobacter pylori PMSS1	*Helicobacter pylori* SS1	*Total*
Wastewater	**3,012**	**8,417**	**11,429**
Wastewater Augsburg	1,439	1,540	2,979
Wastewater Munich	932	2,782	3,714
Wastewater Innsbruck	641	4,095	4,736
Wastewater Westendorf	1,422	n/a	
Stool	**4,958**	**4,522**	**9,480**
Stool sample 002	2,157	2,038	4,195
Stool sample 006	1,536	847	2,383
Stool sample 007	1,265	1,637	2,902
	7,970*	**12,939***	**20,909*(22,331)**

*calculated without Westendorf

As a next step, bacteriophages would be mapped to reference databases, but bacteriophages do not have a marker gene, which would help them identify and categorize them. As a consequence, reference databanks are incomplete and filled with partial information. Therefore, an in-house *Helicobacter pylori* phage database was created. For that, 389 *Helicobacter pylori* sequences were downloaded from NCBI and screened for prophage regions. 90 prophage regions were predicted and put into the database together with 41 *Helicobacter* phage sequences, which were already available from NCBI. Together, they made 131 phage sequences in the in-house HP databank.

To see which phages I had isolated for both strains, potential contigs were placed in three categories: Cat1, Cat2 and Cat3. Viral contigs were put into the first category when iPHoP predicted *Helicobacter pylori* as their host. So, Cat1 is based on host prediction. Cat2 was for all contigs where no host prediction worked; no matches could be found in the NT prokaryotic or NT virus database and the in-house HP database. Cat2 was used for potential novel phages. If contigs could be mapped to the in-house HP database, they were classified as Cat3.

For *Helicobacter pylori* PMSS1, a total of 2,780 were classified as Cat1 with no overlap to the 6,276 contigs placed in Cat2. In Cat3, only one contig was put since 335 shared Cat3 with Cat1. So, 335 contigs could be found in the in-house databank and *Helicobacter pylori* was predicted as their host. 3,269 contigs were placed in Cat1 for the SS1 strain. 9,297 contigs could not be mapped to any sequences in the in-house database and, had no host prediction and were therefore put into Cat2. Two contigs could be found in the in-house HP database and 371 contigs were found in the database and had *H. pylori* predicted as their host. There was no overlap for the categories one and two, two and three, and none for all three, neither for *H. pylori* PMSS1 nor for SS1. The distribution and overlaps can be seen in Figure 28.

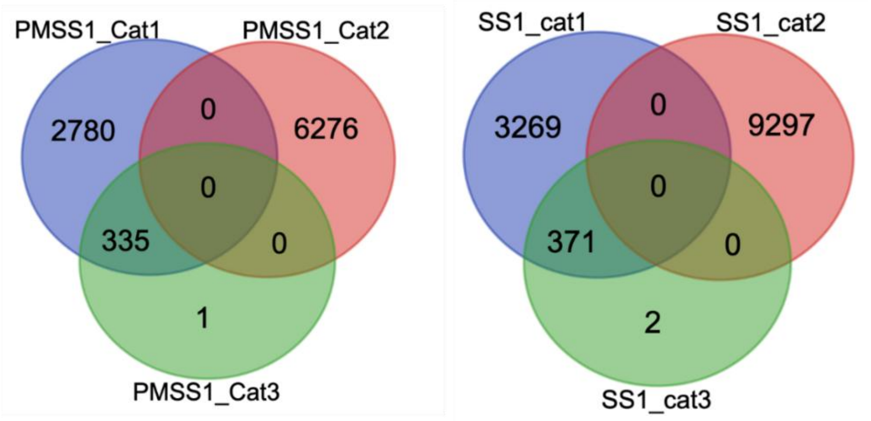


Figure 28_Categorical Distribution of Contigs

Contigs that were assembled from viral-tagged cells isolated with PMSS1 are seen on the left side, whereas on the right, contigs are shown, which were found with SS1. None of the contigs were found in all three categories or in Cat1 and Cat2. Overlaps are between Cat1 and Cat3. Cat2 has for both strains the most found contigs whereas Cat3 for both the least.

To refine the analysis, only contigs longer than 10 kbp were considered, which left us with 599 contigs. From those, 599,192 contigs belong to the *H. pylori* PMSS1, and 407 to *H. pylori* SS1. Further investigation showed that from those 192 contigs for PMSS1, 11 came from stool sample 002, 93 from stool sample 006, and 85 from stool sample 007. Contigs that were longer than 10 kbp were not found in any wastewater (Munich, Augsburg, Innsbruck) except for three contigs coming from Westendorf. As Westendorf wastewater was only tested on the PMSS1 strain, we considered that no phage was found in any wastewater source for that strain. A similar picture was seen for the stool sample for SS1; with 27, the least amount of contigs was calculated for 002, then the highest amount was found for 006 with 83, and 007 was in the middle with 46. A completely different situation was seen for wastewater contigs. Compared to PMSS1, SS1 had 36 contigs from Augsburg, 112 from Innsbruck, and 103 from Munich. When we categorized all found contigs from PMSS1, 191 were placed in Cat1 and one in Cat2. For *H. pylori* SS1, 374 contigs were classified as Cat1 and 33 as Cat2. To even further refine the outcome, all redundant contigs were removed, and a final number of 491 contigs were retained. Those contigs were then used to create a pseudo-phylogenetic tree via pairwise Average

Nucleotide Identity (Figure 29, Supplementary Figure 3). The tree consists of the 491 non-redundant contigs found from the viral tagging (green), then 90 prophage predictions from *Helicobacter pylori* genomes (pink), 41 sequences from *Helicobacter pylori* phages (red) present in the public database, one Coronavirus sequence (yellow), two *Helicobacter pylori* PMSS1 bacterial chromosome and plasmid sequences (bright and dark grey), one chromosome and plasmid sequence from *Helicobacter pylori* strain SS1 (bright and dark blue) and one virome sample (light orange).

The tree shows that publicly available sequences group together apart from two prophage clusters, which seem to have a closer relation to viral-tagged contigs. The viral-tagged contigs separate from the in-house database sequences as branches are rarely shared or connected.

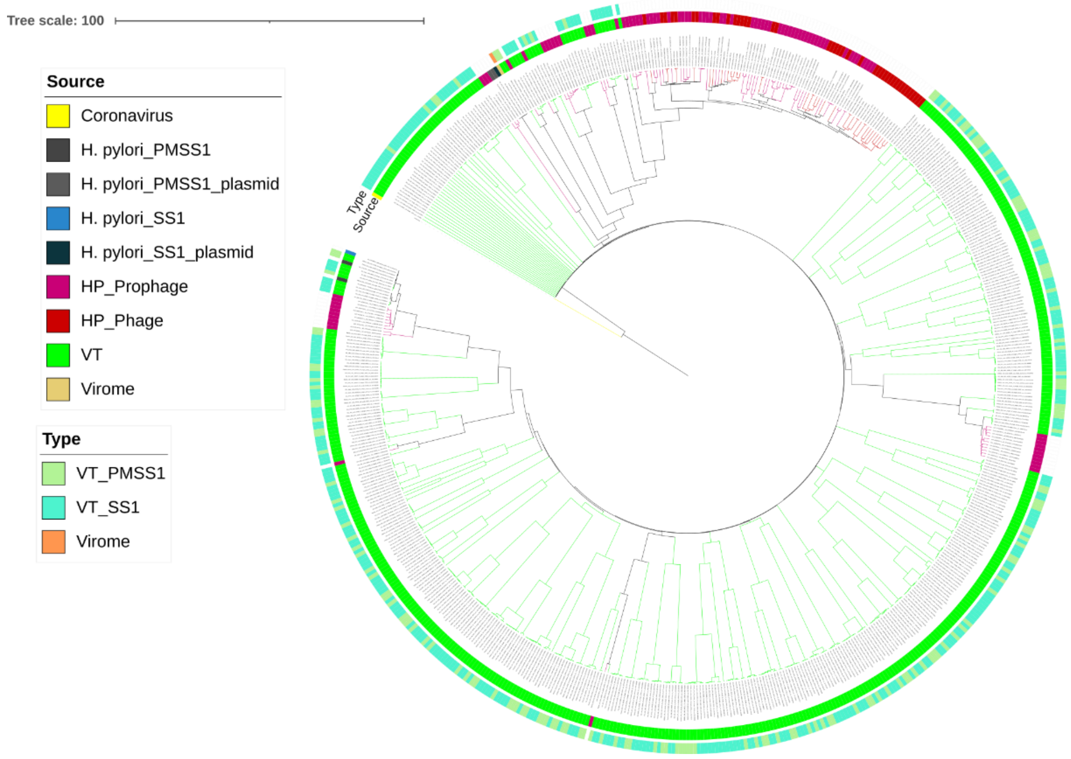


Figure 29_Phylogenetic Tree Viral-Tagged Contigs

The phylogenetic tree shows the relationship between publicly available sequences (red and purple), viral-tagged contigs (green), bacterial genomes (grey and blue), and Coronavirus (yellow). The outer ring differentiates between viral-tagged contigs isolated with SS1 (light blue) or PMSS1 (light green).

We added a second ring to differentiate between the two viral-tagged strains, *H. pylori* PMSS1 and SS1. We also included the virome sample that overlapped with the Source ring's virome. Contigs that belong to the strain SS1 are presented in bright blue and in light green for PMSS1. The tree shows no significant separation between contigs, which were found with PMSS1 or with SS1. Contigs isolated with SS1 do share relatives with contigs isolated from PMSS1. This result was also seen by the heatmap where contigs were placed on the x- and y-axis and compared with their ANI (color-code). There is no clustering that contigs isolated with SS1 would be more similar to each other than contigs isolated with PMSS1. On the contrary, the heatmap has various small to bigger dark areas representing contigs, which have nothing in common with the compared contigs. The heatmap (Supplementary Figure 4) shows that all found contigs are highly diverse with higher dissimilarity on their ANI apart from 100 % matches when contigs were compared with themselves.

The same picture is drawn in those PCA plots (Figure 30). The first PCA plot compares Category 1 with Category 2 isolated either on PMSS1 or SS1. The comparison shows that there is no clustering of contigs isolated on the two strains. Some are more similar to each other than others. Contigs found in Category 2 are less spread than in Category 1 and underrepresented.

When we compared multiple sources and datasets, we got a clear distribution between publicly available virulent *H. pylori* phages and *H. pylori* contigs isolated with viral tagging. As already seen in the phylogenetic tree, viral tagging contigs are highly unrelated to those from the public dataset. Phages from the public database cluster together, whereas a few prophages are more spread than virulent phages. Three prophages are separated from all publicly available sequences and from all contigs found via viral tagging. A small set of prophages are more similar to viral-tagged found contigs than to publicly available *H. pylori* phages. Those prophages cluster with contigs found using *H. pylori* SS1. One prophage overlaps with one contig from Category 2 isolated with *H. pylori* PMSS1.

Despite this huge separation between publicly available *H. pylori* phages and isolated ones. There is one contig isolated with *H. pylori* PMSSS1 positioned in the middle between those two clusters.

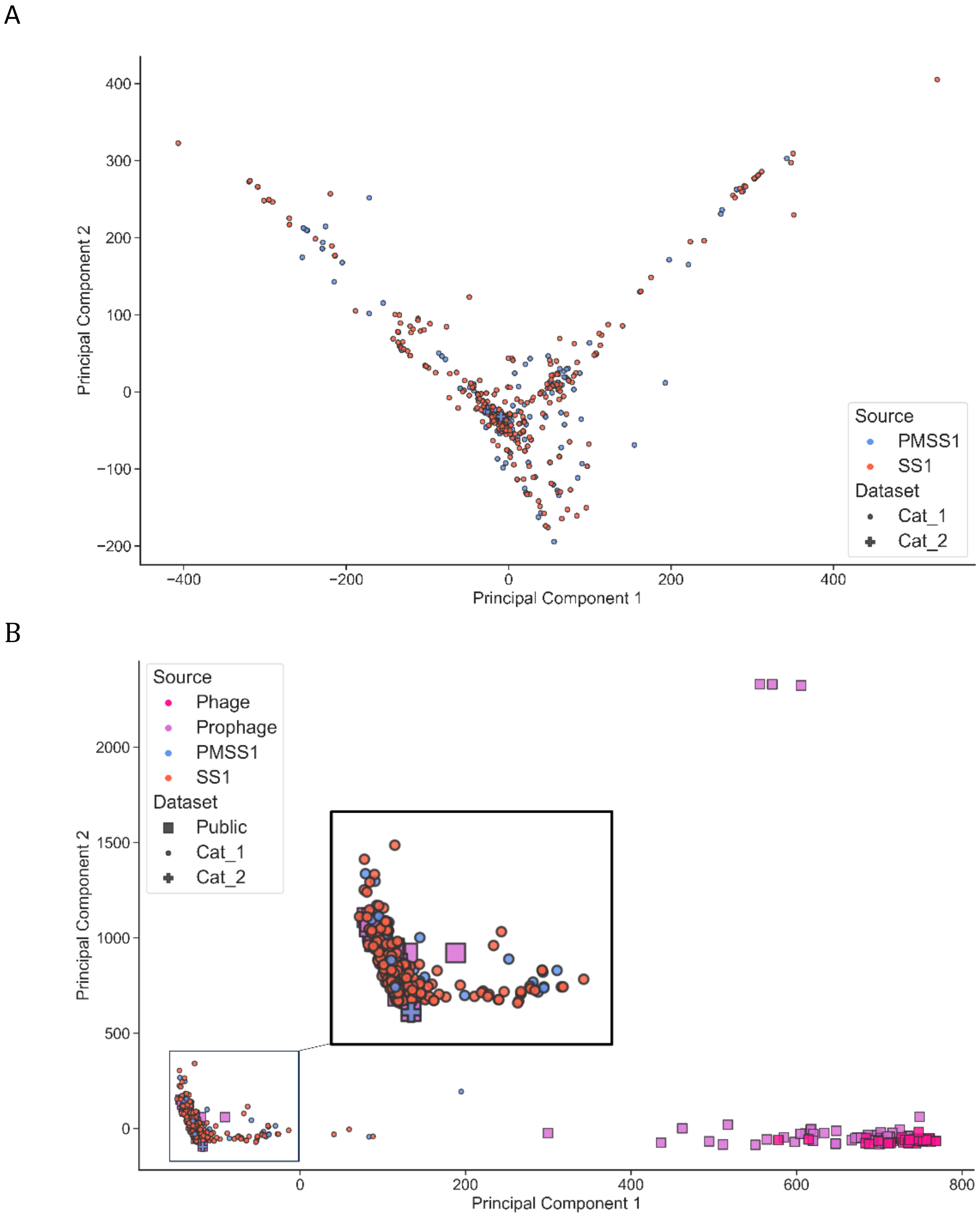


Figure 30_PCA Plots Comparison Between Viral-Tagged Contigs and Public Database

A_PCA plot shows the similarity of contigs isolated with SS1 or PMSS1 with the assigned category.

B_PCA plot visualizes the similarity and dissimilarity between phages and prophages coming from public databases and contigs from viral tagging.

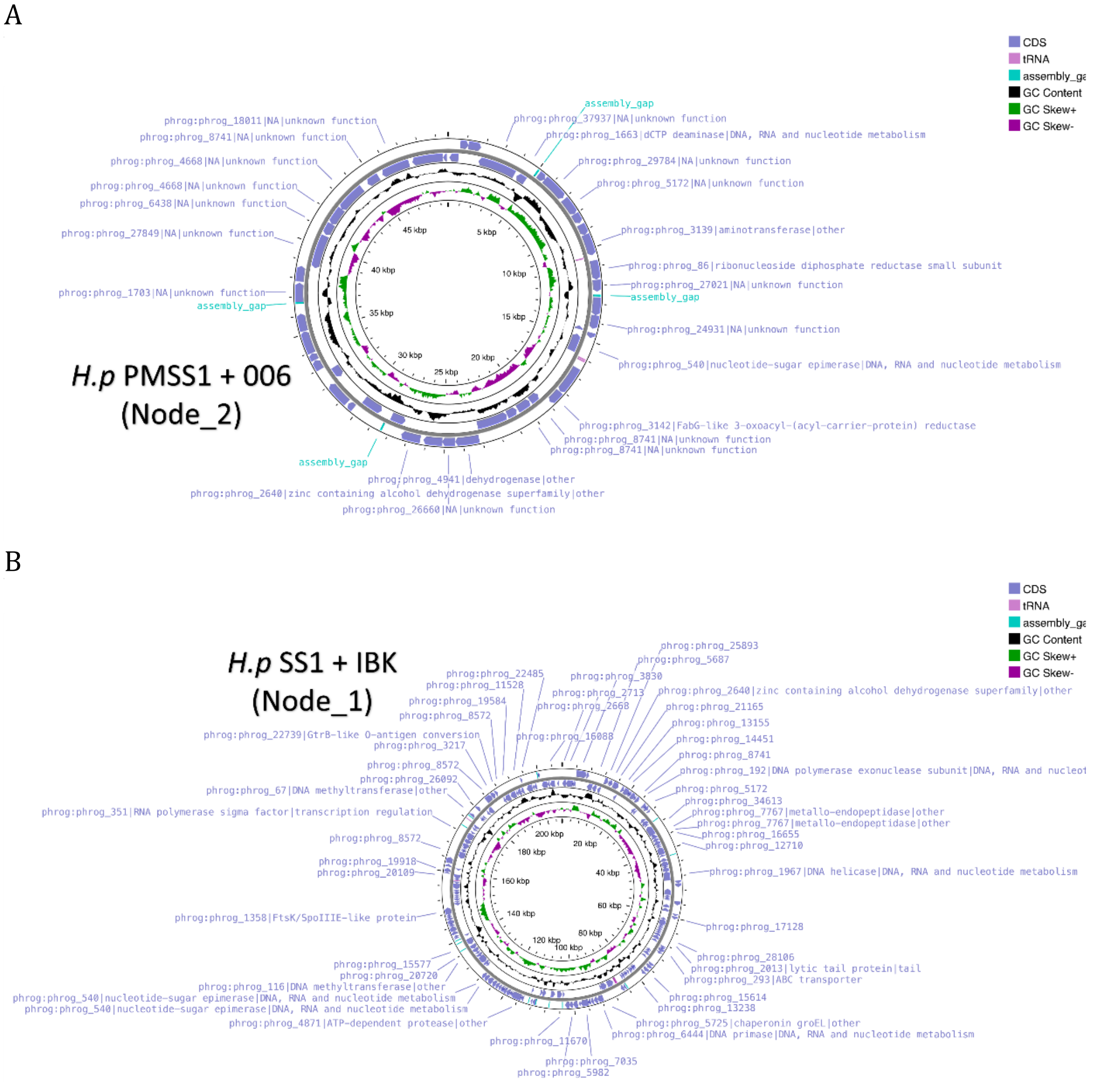
Since the strains do not separate the contigs from each other and the contigs themselves are so diverse, we wanted to know how many similar contigs were found but isolated on the other *H. pylori* strains. In total, eight contigs were detected. In cluster 56, the contig isolated from the mix *H. pylori* SS1 with Innsbruck wastewater has a 92.56 % similarity to a contig isolated on *H. pylori* PMSS1 with the stool sample 006. In cluster 152, the isolated contig from PMSS1 with the stool sample 007 matched 94.53 % with the contig found with SS1 and Munich wastewater. A 99.98 % overlap was found for the matches PMSS1 and stool 002 with SS1 and Augsburg wastewater. A total of 100 % matches were found for clusters 208, 442, and 322, which all involved PMSS1 with the stool sample 007 and then SS1 with Innsbruck wastewater or SS1 with the same stool sample, respectively, for the latter two clusters. Cluster 331 mapped the contig found with SS1 and 006 stool to PMSS1 and stool number 007 with 98.03 %. In cluster 346, a contig from SS1 with Munich wastewater was matched to 91.35 % to a contig coming from PMSS1 and stool sample 006. The strain with the asterisk is the reference strain. The list with identical contigs can be seen in Table 4.

*Table 4_List of Identical Phages Found with H. pylori PMSS1 and with H. pylori SS1 (*reference sequence)*

Cluster	Contigs	% of Similarity
Cluster 56	>SS1_Ib4_23L004816_S9_L001_NODE_22_length_33965* >PMSS1_006_23L004809_S2_L002_NODE_25_length_26146	92.56 %
Cluster 152	>SS1_MUC_23L004814_S7_L001_NODE_45_length_20687* >PMSS1_007_23L004810_S3_L002_NODE_103_length_15833	94.53 %
Cluster 195	>PMSS1_002_23L004808_S1_L001_NODE_69_length_18440* >SS1_Augs_23L004815_S8_L002_NODE_38_length_14430	99.98 %
Cluster 208	>PMSS1_007_23L004810_S3_L002_NODE_86_length_17565* >SS1_Ib4_23L004816_S9_L002_NODE_116_length_13282	100 %
Cluster 331	>SS1_006_23L004812_S5_L001_NODE_46_length_13097* >PMSS1_007_23L004810_S3_L002_NODE_177_length_12089	98.03 %
Cluster 346	>SS1_MUC_23L004814_S7_L001_NODE_149_length_12663* >PMSS1_006_23L004809_S2_L001_NODE_213_length_11105	91.35 %
Cluster 443	>PMSS1_007_23L004810_S3_L001_NODE_208_length_10715* >SS1_007_23L004813_S6_L001_NODE_148_length_10200	100 %
Cluster 322	>PMSS1_007_23L004810_S3_L002_NODE_142_length_13179* >SS1_007_23L004813_S6_L002_NODE_156_length_10297	100 %

After cleaning up the data, we found four contigs with high potential. Those contigs did match *H. pylori* as host predictions and were therefore classified as Category 1. Their length is longer than 10 kbp, and genome quality was rated as medium. The first contig was found with *H. pylori* PMSS1 and the stool sample 006. The contig has 49,627 base pairs (Figure 31 A). Gene prediction identified enzymes for RNA and DNA metabolism as well as a nucleotide-sugar epimerase, a reductase, or dehydrogenases, but for the majority of the genome, their protein functions are mainly unknown.

The other three highly potential contigs were all found with *H. pylori* SS1 and Innsbruck wastewater. There was one longer contig with 206,764 basepairs and two smaller contigs with 25,993 and 23,019 basepairs (Figure 31 B-D). Most predicted proteins belong either to DNA/RNA metabolism or DNA methyltransferases, transcription, or elongation factors. However, most proteins were classified as unknown or not identifiable as such.


A
CDS
tRNA
assembly_ga
GC Content
GC Skew+
GC Skew-
assembly_gap
phrog:phrog_18011|NA|unknown function
phrog:phrog_8741|NA|unknown function
phrog:phrog_4668|NA|unknown function
phrog:phrog_4668|NA|unknown function
phrog:phrog_6438|NA|unknown function
phrog:phrog_27849|NA|unknown function
phrog:phrog_1703|NA|unknown function
assembly_gap
phrog:phrog_37937|NA|unknown function
phrog:phrog_1663|dCTP deaminase|DNA, RNA and nucleotide metabolism
phrog:phrog_29784|NA|unknown function
phrog:phrog_5172|NA|unknown function
phrog:phrog_3139|aminotransferase|other
phrog:phrog_86|ribonucleoside diphosphate reductase small subunit
phrog:phrog_27021|NA|unknown function
assembly_gap
phrog:phrog_24931|NA|unknown function
phrog:phrog_540|nucleotide-sugar epimerase|DNA, RNA and nucleotide metabolism
phrog:phrog_3142|FabG-like 3-oxoacyl-(acyl-carrier-protein) reductase
phrog:phrog_8741|NA|unknown function
phrog:phrog_8741|NA|unknown function
assembly_gap
phrog:phrog_4941|dehydrogenase|other
phrog:phrog_2640|zinc containing alcohol dehydrogenase superfamily|other
phrog:phrog_26660|NA|unknown function
45 kbp
5 kbp
40 kbp
10 kbp
35 kbp
15 kbp
30 kbp
20 kbp
25 kbp
H.p PMSS1 + 006
(Node_2)
B
CDS
tRNA
assembly_ga
GC Content
GC Skew+
GC Skew-
H.p SS1 + IBK
(Node_1)
phrog:phrog_25893
phrog:phrog_5687
phrog:phrog_22485
phrog:phrog_11528
phrog:phrog_3830
phrog:phrog_2713
phrog:phrog_19584
phrog:phrog_2668
phrog:phrog_8572
phrog:phrog_2640|zinc containing alcohol dehydrogenase superfamily|other
phrog:phrog_21165
phrog:phrog_13155
phrog:phrog_22739|GtrB-like O-antigen conversion
phrog:phrog_3217
phrog:phrog_16088
phrog:phrog_14451
phrog:phrog_8741
phrog:phrog_8572
phrog:phrog_26092
phrog:phrog_67|DNA methyltransferase|other
phrog:phrog_192|DNA polymerase exonuclease subunit|DNA, RNA and nucleot
phrog:phrog_5172
phrog:phrog_34613
phrog:phrog_351|RNA polymerase sigma factor|transcription regulation
phrog:phrog_7767|metallo-endopeptidase|other
phrog:phrog_7767|metallo-endopeptidase|other
phrog:phrog_16655
phrog:phrog_12710
phrog:phrog_8572
phrog:phrog_19918
phrog:phrog_20109
phrog:phrog_1967|DNA helicase|DNA, RNA and nucleotide metabolism
phrog:phrog_1358|FtsK/SpoIIIE-like protein
phrog:phrog_17128
phrog:phrog_28106
phrog:phrog_2013|lytic tail protein|tail
phrog:phrog_293|ABC transporter
phrog:phrog_15577
phrog:phrog_20720
phrog:phrog_116|DNA methyltransferase|other
phrog:phrog_540|nucleotide-sugar epimerase|DNA, RNA and nucleotide metabolism
phrog:phrog_540|nucleotide-sugar epimerase|DNA, RNA and nucleotide metabolism
phrog:phrog_15614
phrog:phrog_13238
phrog:phrog_4871|ATP-dependent protease|other
phrog:phrog_5725|chaperonin groEL|other
phrog:phrog_6444|DNA primase|DNA, RNA and nucleotide metabolism
phrog:phrog_11670
phrog:phrog_7035
phrog:phrog_5982
200 kbp
20 kbp
180 kbp
40 kbp
160 kbp
60 kbp
140 kbp
80 kbp
120 kbp
100 kbp

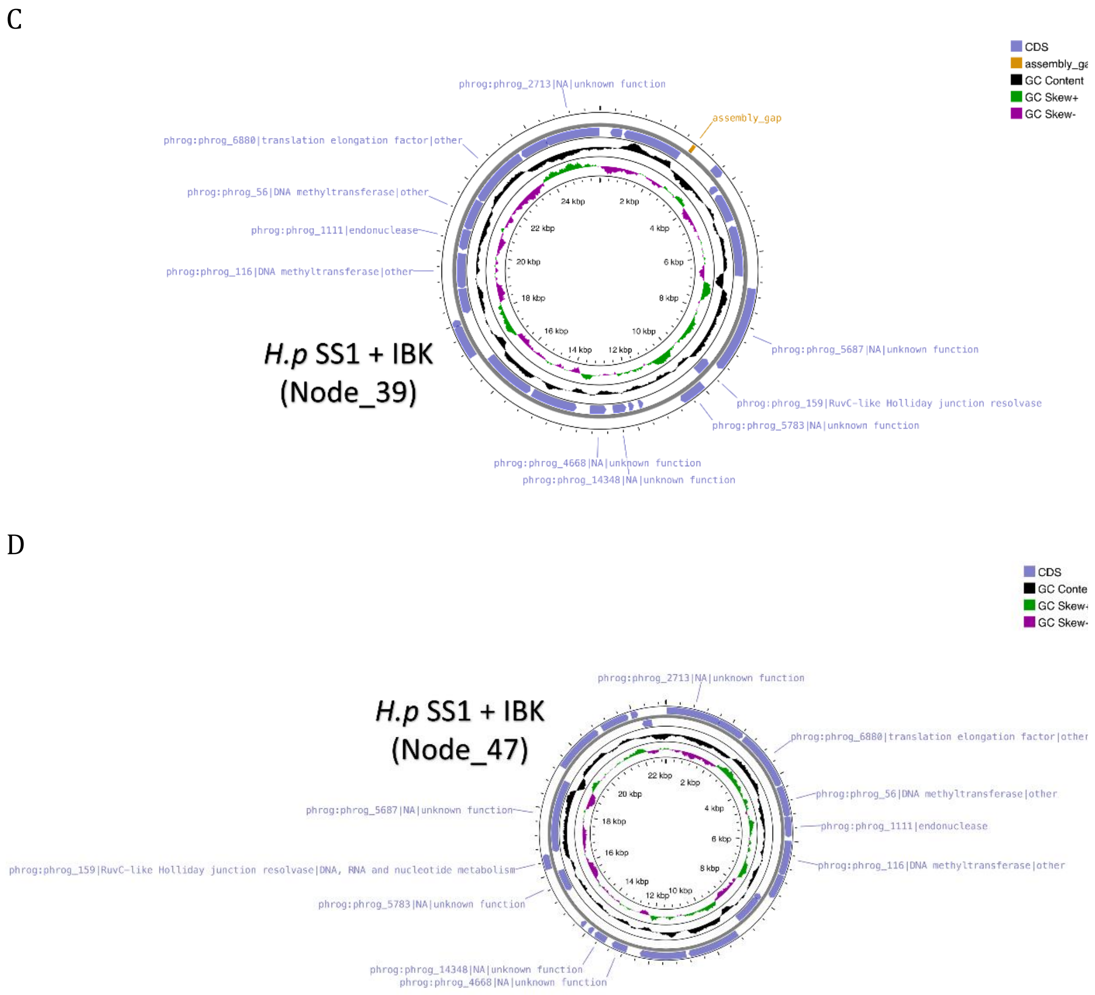

C
CDS
assembly_ga
GC Content
GC Skew+
GC Skew-
phrog:phrog_2713|NA|unknown function
assembly_gap
phrog:phrog_6880|translation elongation factor|other
phrog:phrog_56|DNA methyltransferase|other
phrog:phrog_1111|endonuclease
phrog:phrog_116|DNA methyltransferase|other
2 kbp
4 kbp
6 kbp
8 kbp
10 kbp
12 kbp
14 kbp
16 kbp
18 kbp
20 kbp
22 kbp
24 kbp
H.p SS1 + IBK
(Node_39)
phrog:phrog_5687|NA|unknown function
phrog:phrog_159|RuvC-like Holliday junction resolvase
phrog:phrog_5783|NA|unknown function
phrog:phrog_4668|NA|unknown function
phrog:phrog_14348|NA|unknown function
D
CDS
GC Conte
GC Skew
GC Skew-
phrog:phrog_2713|NA|unknown function
H.p SS1 + IBK
(Node_47)
phrog:phrog_6880|translation elongation factor|other
phrog:phrog_56|DNA methyltransferase|other
phrog:phrog_1111|endonuclease
phrog:phrog_116|DNA methyltransferase|other
phrog:phrog_5687|NA|unknown function
phrog:phrog_159|RuvC-like Holliday junction resolvase|DNA, RNA and nucleotide metabolism
phrog:phrog_5783|NA|unknown function
2 kbp
4 kbp
6 kbp
8 kbp
10 kbp
12 kbp
14 kbp
16 kbp
18 kbp
20 kbp
22 kbp
phrog:phrog_14348|NA|unknown function
phrog:phrog_4668|NA|unknown function

Figure 31_Potential Phages From Category 1

Those four phages have *H. pylori* as host prediction. They are all longer than 10 kbp with medium quality. In addition, some of their genes could be annotated. Picture A shows the phage isolated with *H. pylori* PMSS1 and the stool sample 006. Phages B to D were all isolated with *H. pylori* SS1 and Innsbruck wastewater.

Additionally, to those four highly potential contigs from Category 1, we also checked 34 contigs from Category 2. Those contigs did not match any NT prokaryotic or virus databank or any other database and are all longer than 10 kilobases. Of those 34, we removed all redundant ones, and 25 remained (Figure 32, Supplementary Figure 5). 22 out of those 25 are pairs from *H. pylori* SS1 with the wastewater from Innsbruck. Two came from Munich wastewater and from *H. pylori* SS1 as well. The last one was isolated with the PMSS1 strain and wastewater from Westendorf. Protein comparison showed no similarity at all between those 25 contigs. These contigs have a highly diverse composition. The only small similarity is between contig SS1_Ibk_13 and SS1_Ibk_10.

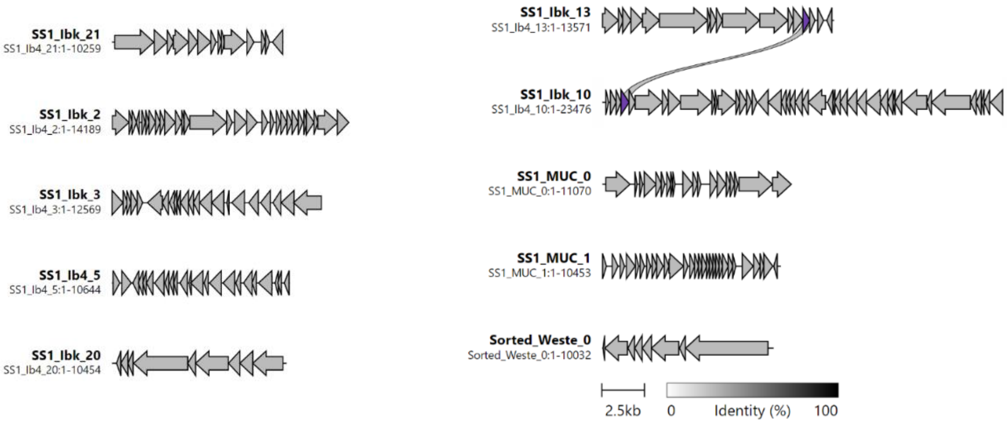


Figure 32_Highly Potential Phages From Category 2

Those contigs in Category 2 did not have *Helicobacter pylori* predicted as their host and did not match the in-house database and sequences longer than 10 kbp. Apart from their shared host, protein comparison showed no similarity apart from one between SS1_Ibk13 and SS1_Ibk10.

After analyzing the most potential contigs further, the results showed a lot of dissimilarity and many unknowns. To shed light on those unknowns, viral-tagged contigs were compared genetically with publicly available sources. We first investigated auxiliary metabolic genes and found that only the nicotinamide adenine dinucleotide (NAD)-dependent epimerase/dehydratase family appears in both sample groups. All other AMGs are either only present in our in-house database or in the viral-tagged contigs as Figure 33 shows.

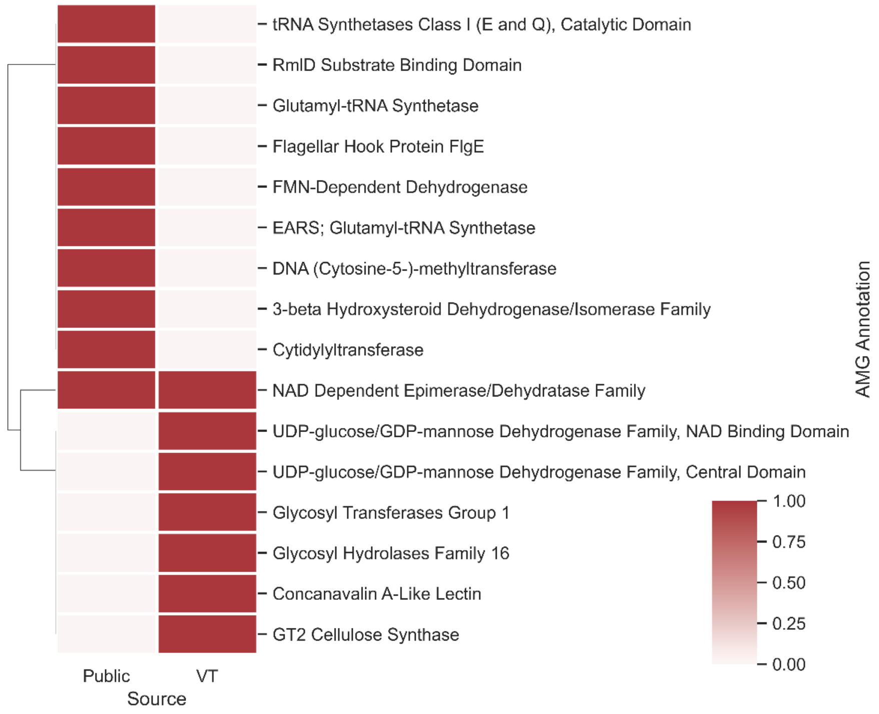


Figure 33_Auxiliary Metabolic Gene Annotation

Our in-house database and viral-tagged samples were screened for AMGs and compared. The result shows that apart from the NAD-dependent epimerase/dehydratase family, all other AMGs are highly different between those two groups.

Since the AMGs only strengthened the high diversity amongst phages, we looked into gene function categories. We had eight categories plus others, which excluded the other eight and everything else. The graphs show an absolute number in bright green viral-tagged isolated contigs and in purple phages from the public database. If overlaps occurred, they were presented in bright blue (Supplementary Figure 6). In five categories, the number of gene clusters was higher in the viral-tagged contigs than in the public phage database. Those five categories are “DNA, RNA and nucleotide metabolism”, “Connector”, “Lysis”, “Moron, auxiliary metabolic gene and host takeover”, “Tail”, and “other”. Viral-tagged contigs also had more gene clusters than the public ones in the category “Head and packaging” as well as in “Transcription regulation”. In the category “Integration and excision” no gene cluster is visible for viral-tagged contigs but there is an overlap. However, in this category, the publicly available sequences have many more gene clusters. When looking into overlaps, three categories have a minor overlap, “Transcription regulation,” “Integration and excision,” and “Head and packaging.” Whereas “DNA, RNA and nucleotide metabolism”, “Moron, auxiliary metabolic gene and host takeover" and "others" have a major overlap. However, these are all in absolute numbers and not really comparable, and therefore, we looked at gene ratios, which presented a totally different picture (Figure 34 A). In six out of nine categories, genes from public databases are more highly represented than those from viral-tagged contigs. Only in "Moron, auxiliary metabolic gene and host takeover" as well as in "others", viral-tagged contigs have a higher percentage of genes than the public dataset. The category "Connector" is underrepresented in both datasets. When looking into the total gene number and gene clusters, we found the following (Figure 34 B). Viral tagging contigs have a total of 12,004 genes compared to 3,478 genes from the public database. Clustering those genes showed that both datasets share 231 gene clusters, and 2,407 clusters were only found in the VT dataset and 249 only in the public counterpart. The gene clusters are based on the most frequent gene annotation and its function.

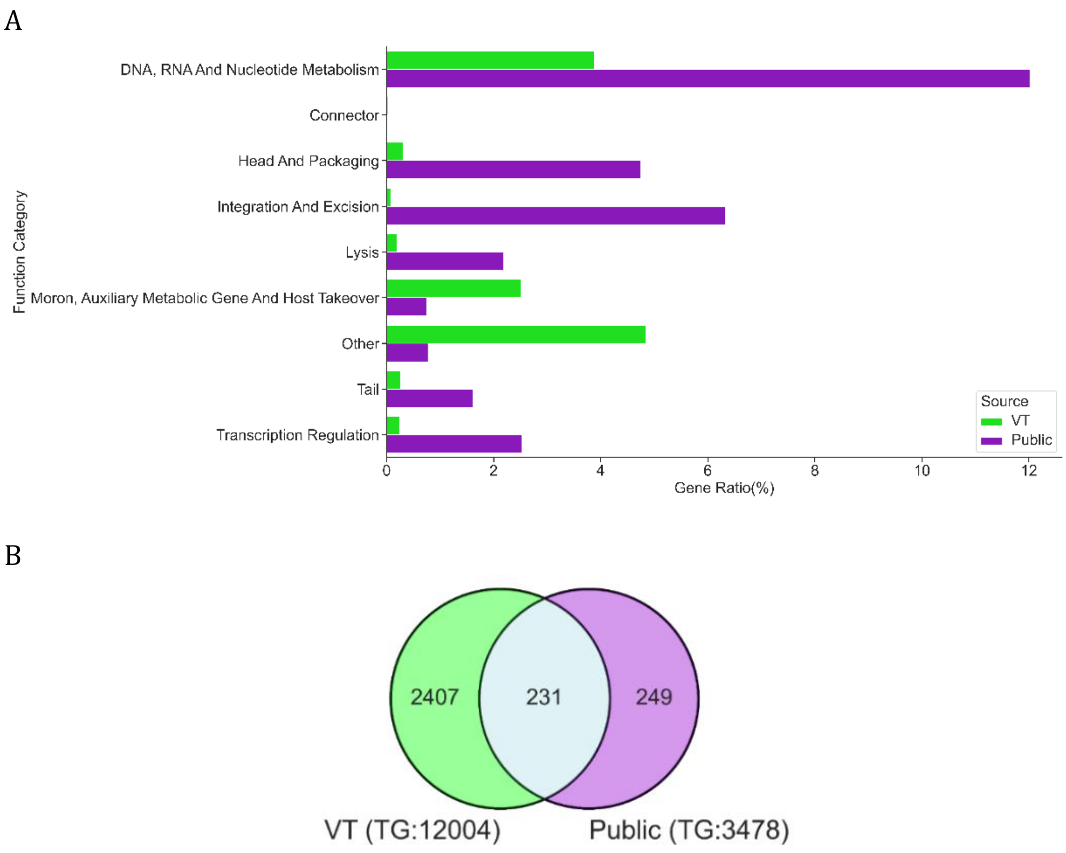


Figure 34_Gene Function Clustering

Figure _A shows the ratio of functional genes for the in-house database and viral-tagged samples.

Figure_B is a veen plot. The plot shows the total amount of genes (VT:12,004 vs. Public:3,478), and how many clusters were found for each group and their overlap.

PHROG (Prokaryotic Virus Remote Homologous Groups) analysis showed that 18 genes are more prevalent in viral-tagged contigs, while 18 other genes were more predominant in phages coming from the in-house databank. Structural genes (head and packaging), transcription regulation, and lysis genes were more abundant in phages from the in-house databank, while DNA, RNA, and nucleotide metabolic genes were shared. ATPase and tRNA-methyltransferase were only detected in the viral-tagged contigs. All gene functions in detail are presented in Figure 35.

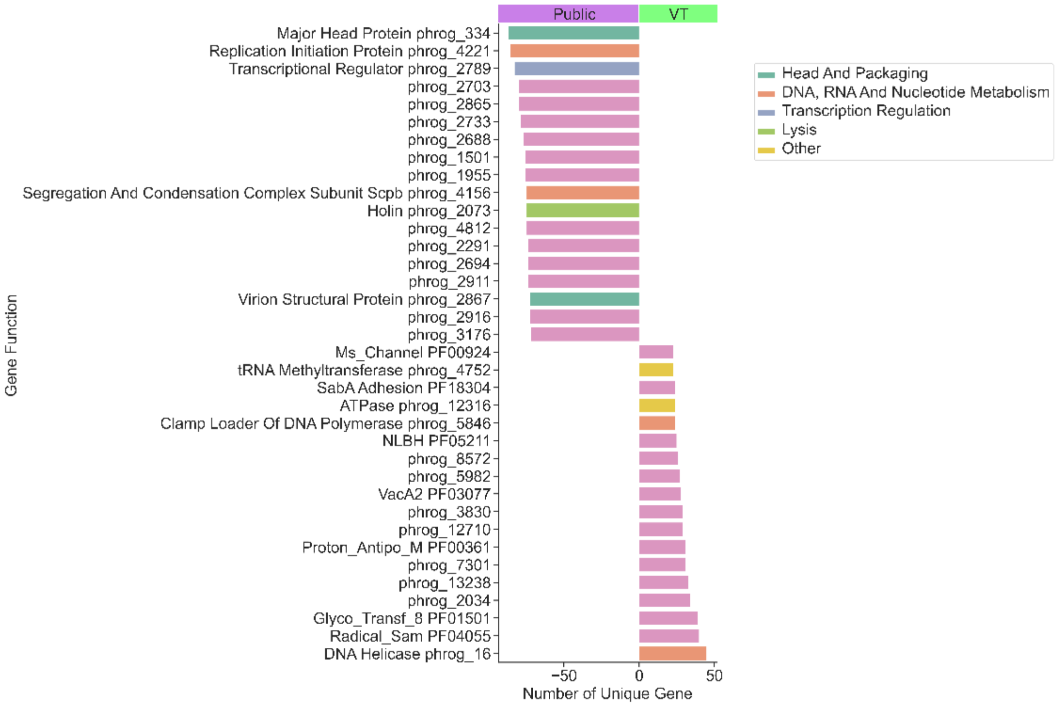


Figure 35_Unique Genes Identified in Viral-Tagged and In-House Database

18 unique genes were found for both groups, but only DNA, RNA, and nucleotide metabolic genes were shared. Others, like structural genes, transcription regulation, or lysis, were only detected in the in-house phages, where ATPases or tRNA-methyltransferases were only seen in viral-tagged contigs.

As the last point, we looked into endolysins, as they, on their own, are a promising antibacterial agent. We screened our in-house database and our 599 sequences from viral tagging for endolysins and compared them to each other. We found 20 endolysins, whereas only one was detected in our in-house databank and 19 in the viral-tagged samples (Figure 36).

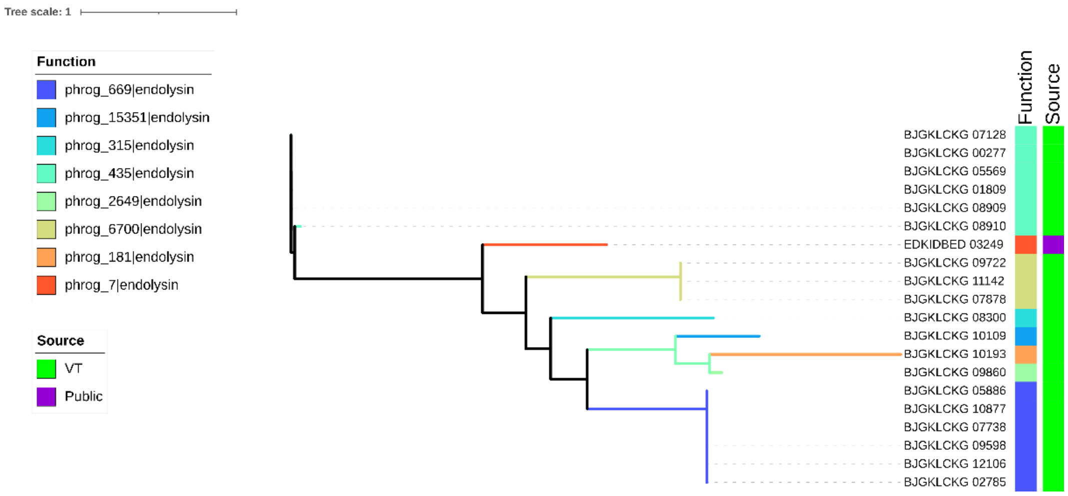


Figure 36_Phylogenetic Tree of Endolysins Found in Viral-Tagged Samples and In-House Databank

Viral-tagged samples are presented in green, whereas the in-house databank sequence is in purple. Only one was detected in the in-house databank and 19 for the viral-tagged contigs. The functions were based on the PHROG database, and eight phrog groups were found in total.

In the PHROG database, all of them were annotated as endolysins and placed in the functional category: lysis. The amount of protein sequences of which the detected phrog is made ranges from 6 sequences up to 3,145 sequences (Table 5). The latest belongs to phrog_7 which was found to be the only endolysin from the in-house databank. Despite being the biggest phrog and the only one from the in-house databank, the phylogenetic tree shows that phrog_7 is closer related to all others than phrog_435, which is the most distant one. The closest relation is between phrog_181 and phrog_2649, followed by phrog_15251. Six endolysins clustered together in phrog_669, the same as in phrog_435. In phrog_6700, three endolysins were detected.

Table 5_Description of phrogs

phrog_Number	PHROG Protein Sequences	Annotated as / Functional Category	PFAM Prediction
phrog_669	210	Endolysin / Lysis	Transglycosylase SLT Domain
phrog_15351	6	Endolysin / Lysis	D-Alanyl-D-Alanine Carboxypeptidase
phrog_315	389	Endolysin / Lysis	Glycosyl Hydrolase
phrog_435	299	Endolysin / Lysis	N-Acetylmuramoyl-L-Alanine Amidase
phrog_2649	54	Endolysin / Lysis	D-ala-D-ala Dipeptidase
phrog_6700	18	Endolysin / Lysis	Mannosyl-Glycoprotein endo-beta-N-Acetylglucosaminidase
phrog_181	576	Endolysin / Lysis	Chitinase Class I
phrog_7	3,145	Endolysin / Lysis	Phage Lysozyme

3.3 Colorectal Cancer Cross-Infection Study

3.3.1 Bacteriome and Virome

Deciphering the complex interactions between the gut microbiota and their host is crucial for a better understanding of diseases in the future. The more we understand the connections and consequences, the more we can help patients.

Here, we want to better understand the role of phages within their gut community under certain conditions. We examined three stool samples from patients with ulcerative colitis, three samples from early colorectal cancer patients, and two samples from advanced colorectal cancer. All samples were compared to three healthy individuals.

We first looked at the bacterial composition and the relative abundance. We discovered that healthy individuals differ from patient samples. The stool samples from healthy individuals are dominated with *Escherichia spp* (48.26 %) and include small portions of *Bacteroides* (6.76 %) and *Kluyvera*. Traces of *Elizabethkingia*, *Leuconostoc*, *Streptococcus*, and *Pseudoflavonifractor* are detected, but the remaining 37 % are unknown. The unknown proportion increases in all disease conditions: 46 % for advanced cancer, 57 % for early cancer, and the highest with 64 % for the Ulcerative Colitis sample. A significant (p-value <0.01) decrease was also seen for the *Escherichia* genus in all three disease conditions. Ulcerative Colitis had the least amount with 1.07 %, followed by early cancer with 3.28 % and advanced cancer with 6.47 %. An opposite picture was presented for *Bacteroides*, *Clostridium*, and some others. *Bacteroides* were the least abundant genus in the healthy sample. The abundance increased significantly (p-value <0.01) to 16.28 % in UC samples, 21.77 % in CRCE, and 29.61 % in CRCA. *Leuconostoc* had the same trend with the least amount in Ulcerative Colitis, then higher levels in early cancer, and the highest amount in advanced cancer. It was the same for *Elizabethkingia* and *Pseudoflavonifractor*, *Streptococcus*, and *Bacillus*, which were more abundant in UC samples than they were in early or advanced cancer. *Anaerostipes* was only present in Ulcerative Colitis. Details are visualized in Figure 37.

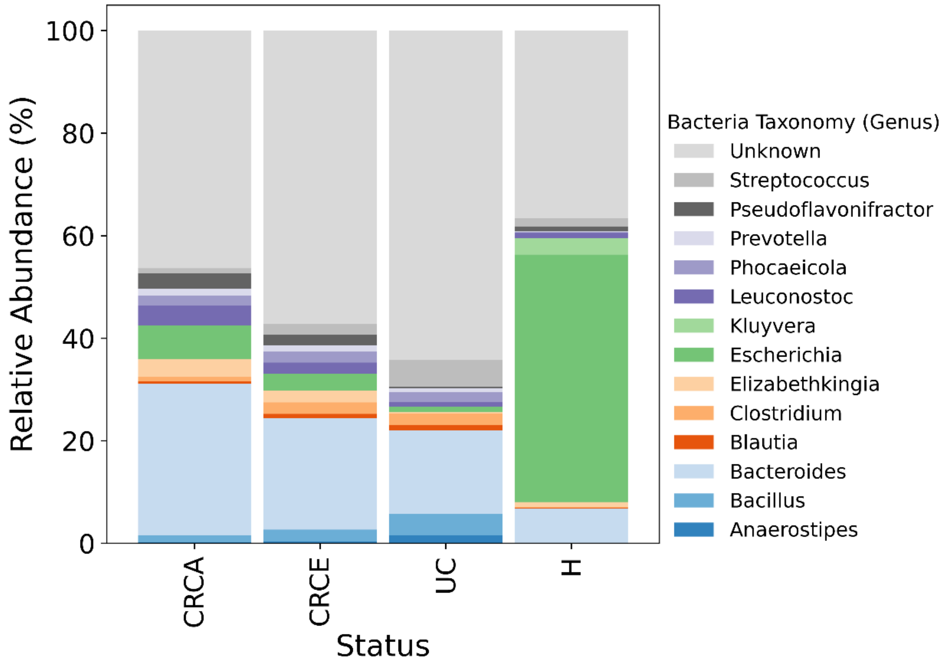


Figure 37_Metagenomic Analysis of Bacterial Samples

The plot shows the relative abundance of bacteria on the genus level for the metagenomic analysis of bacterial samples. Samples from each condition were merged into one category, resulting in CRC_advance (CRCA), CRC_early (CRCE), UC (Ulcerative Colitis), and Healthy (H).

The same analysis has been executed for the virome with very limited information, as the predicted taxonomy was either *Caudoviricetes* or *Malgrandaviricetes*. When conditions were analyzed based on viral clusters, an UpSet plot presented a more differentiated picture (Figure 38).

Viral clusters (VC) were determined, and 140 VCs were detected in CRC advanced, 155 VCs in healthy, 159 VCs in Ulcerative Colitis, and 182 VCs in CRC early. The majority of viral clusters were unique to each condition, and only a few were shared. Nine VCs were detected in CRC early and Ulcerative Colitis. Six were found in CRC early and healthy and 16 were identified in CRC early and advanced. Also, 16 viral clusters were seen between Ulcerative Colitis and healthy. Another 11 were shared between Ulcerative Colitis and CRC advanced. CRC advanced also had 14 viral clusters, which were also detected in healthy samples. Seven were found in healthy, Ulcerative Colitis and early cancer. 12 viral clusters were shared between all disease conditions and only 3 between both cancer samples and healthy. Six were found in advanced cancer, healthy, and Ulcerative Colitis, and only seven were shared between all four conditions.

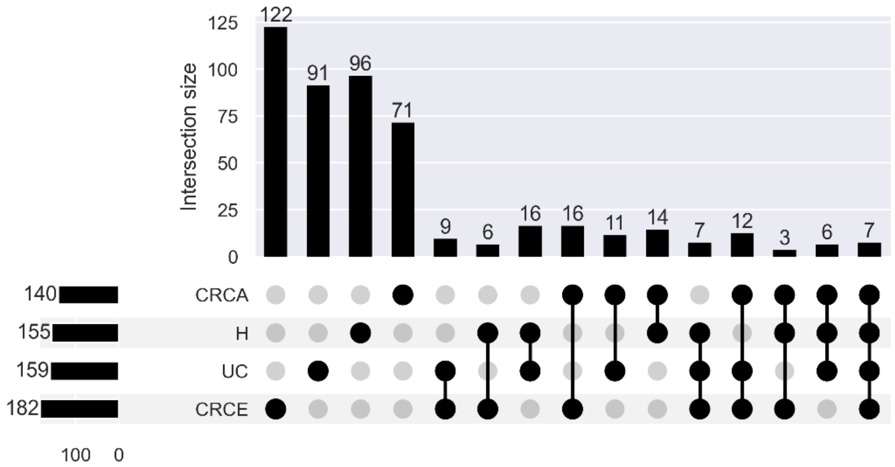


Figure 38_Viral Cluster Distribution in all Four Conditions

The UpSet plot showed that CRCA has the least amount of viral clusters and the CRCE the most. The majority of viral clusters are unique for each condition, and only a few were detected in two or more conditions.

CRCA___Colorectal Cancer Advanced Stadium

CRCE___Colorectal Cancer Early Stadium

UC___Ulcerative Colitis Patients

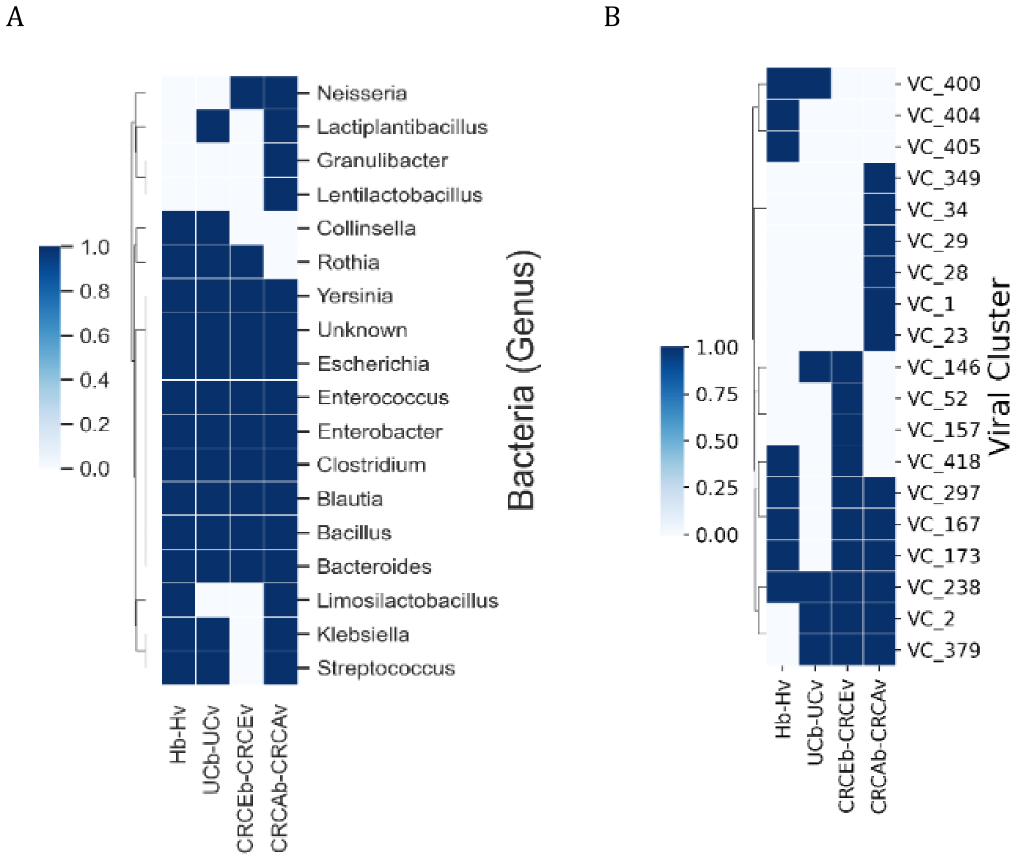
H___Healthy Individuals

3.3.2 Bacterial and Viral Abundance in Cross-Infection

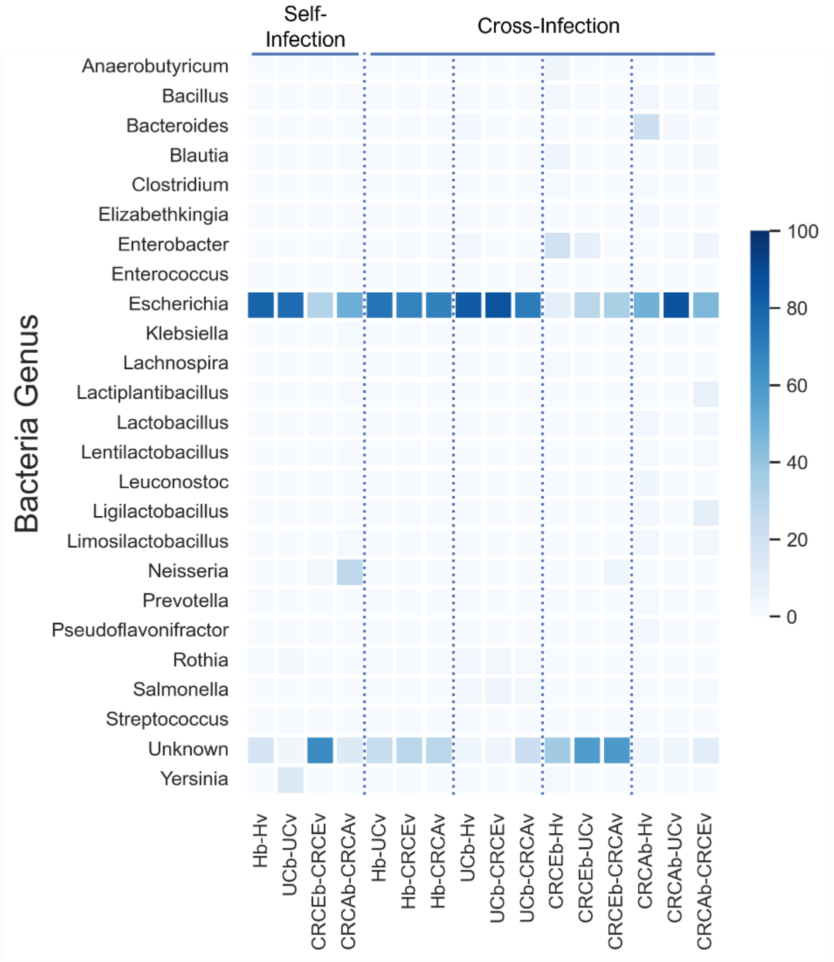
Figure 39 describes the bacterial and viral composition of all cross-infected samples. In Figure 39 A bacterial genera were determined for self-infection samples (bacteria infected with the phages from the same sample).

A big cluster of bacteria is shared in each condition. Those bacterial genera are *Yersinia*, *Escherichia*, *Enterococcus*, *Enterobacter*, *Clostridium*, *Blautia*, *Bacillus*, *Bacteroides*, and unknown. In contrast, only advanced cancer has bacterial genera, which are only present in advanced cancer samples. *Granulibacter* and *Lentibactobacillus* are only identified in advanced cancer. All remaining bacterial genera are shared at least in two conditions. *Neisseria* is present in both cancer samples but not in Ulcerative Colitis or healthy samples. *Neisseria* is also the only one shared by both cancers. *Lactiplantibacillus* is abundant in advanced cancer and Ulcerative Colitis. Ulcerative Colitis shared genera *Streptococcus* and *Klebsiella* with healthy and advanced cancer samples. In advanced cancer, *Limosilactobacillus* was found the same as in healthy samples. Advanced cancer had no trace of *Collinsella* and *Rothia*. *Collinsella* was found in Ulcerative Colitis and healthy samples and *Rothia* in healthy, Ulcerative Colitis and early cancer.

The relative abundance was also compared to better understand the bacterial composition of all cross-infections. In Figure 39 C self-infections were aligned on the left and on the right, all the other cross-infection samples. *Escherichia* was the dominant genus in this heatmap, followed by unknown. CRCAb-Hv had a higher abundance of *Bacteroides* than all the other cross-infections. *Enterobacter* was detected in a higher abundance in CRCEb-Hv and CRCEb-UCv. Also, a higher abundance was seen in *Neisseria* for CRCAb-CRCAv and for *Yersinia* in UCb-UCv.



C


Self-
Infection
Cross-Infection
Bacteria Genus
Anaerobutyricum
Bacillus
Bacteroides
Blautia
Clostridium
Elizabethkingia
Enterobacter
Enterococcus
Escherichia
Klebsiella
Lachnospira
Lactiplantibacillus
Lactobacillus
Lentilactobacillus
Leuconostoc
Ligilactobacillus
Limosilactobacillus
Neisseria
Prevotella
Pseudoflavonifractor
Rothia
Salmonella
Streptococcus
Unknown
Yersinia
Hb-Hv
UCb-UCv
CRCEb-CRCEv
CRCAb-CRCAv
Hb-UCv
Hb-CRCEv
Hb-CRCAv
UCb-Hv
UCb-CRCEv
UCb-CRCAv
CRCEb-Hv
CRCEb-UCv
CRCEb-CRCAv
CRCAb-Hv
CRCAb-UCv
CRCAb-CRCEv
100
80
60
40
20
0

D

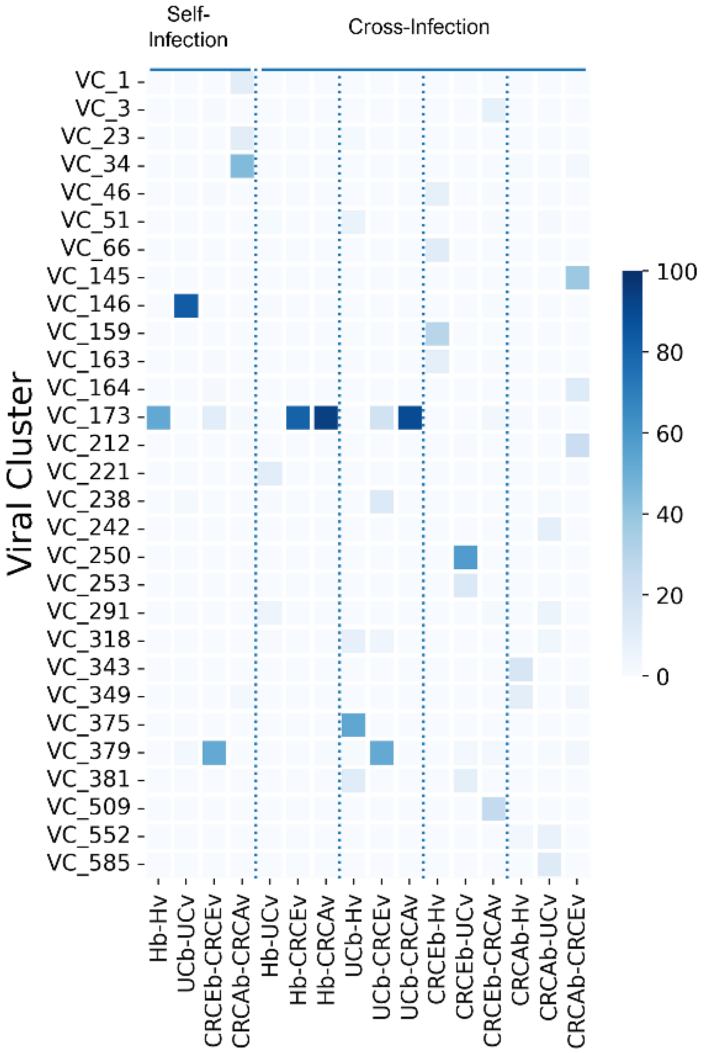


Figure 39_Bacterial Abundance and Viral Cluster Abundance in Cross-Infection Samples

Figures A and B show which bacterial genera and which viral clusters were present. Bacterial genera were shared more than viral clusters, which were more unique for each cross-infection.

Figure C shows the relative abundance of bacterial genera with *Escherichia* being the most dominant one.

Figure D shows the relative abundance of viral clusters. VC173 was present in multiple samples in high abundance. *Escherichia* was predicted as the host of VC173. The remaining high abundance clusters were unique for the samples.

Legend:

H__Healthy Individuals

UC__Ulcerative Colitis Patients

CRCE__Colorectal Cancer Early Stadium

CRCA__Colorectal Caner Advanced Stadium

b__bacteria

v__virus-like-particles

Figure 39 B and D describe the viral composition of cross-infections. Viral contigs were compared and plotted in Figure C to see which ones are present in self-infection samples. In comparison with the bacterial composition, there is only one viral cluster that is shared in all four conditions, viral cluster 238 which host prediction determined *Escherichia*. Yet, ten viral clusters were only identified in one condition and were not shared at all. VC404 and VC405 viral clusters were only detected in healthy samples. The host prediction identified only VC405 as *Escherichia,* VC404 is unknown. Six viral clusters were only seen in advanced cancer. Those were VC349, VC34, VC29, VC28, VC1 and VC23. For the last three, no host could have been predicted; for VC34 and VC349, *Lactobacillus* was predicted as their host, and for VC29, it was *Limosilactobacillus*. Two were identified in early cancer, VC52 and VC157, with *Salmonella* as their predicted host. Ulcerative Colitis has only shared viral clusters. VC400 was detected in healthy and UC, but no host could have been predicted. VC146 was found in UC and early cancer with *Klebsiella* as the predicted host. VC2 and VC379 were detected in UC and both cancer samples. VC379 had *Isoptericola* as its predicted host and VC2 *Escherichia*. VC418 (host unknown) was also shared between healthy and early cancer, and VC297, VC167, and VC173 were detected in healthy, early, and advanced cancer. VC297 had *Raoultella* as the predicted host, and *Escherichia* was predicted for VC167 and VC173.

When relative abundance was compared, there was only one viral cluster, VC173, that was dominant. Hb-Hv (52.50 %) had the highest abundance in the self-infection group, whereas Hb-CRCAv (94.04 %), Hb-CRCEv (80.72 %), and UCb-CRCAv (89.60 %) had the highest among the other cross-infections. VC173 was identified as an *Escherichia* phage based on host prediction. For the remaining viral clusters, a few were identified with a higher abundance. In the self-infection group, there was VC34 in CRCAb-CRCAv (44.70 %), VC146 in UCb-UCv (83.16 %), and VC379 in CRCEb-CRCEv (52.25 %). The predicted hosts were the following: VC34 – *Lactobacillus*, VC146 – *Klebsiella*, VC379- *Isoptericola*. Viral cluster 379 was also abundant in UCb-CRCEv (52 %). CRCAb-CRCEv had the highest abundance in VC145 (38.17 %) and VC212 (22.5 %). VC 145 had *Ligilactobacillus* as host prediction and VC212 *Salmonella*. VC159 was identified with the highest abundance in CRCEb-Hv (29.43 %) but has no host prediction. CRCEb-UCv had viral cluster 250 (57.83 %) detected and VC509 in CRCE-CRCAv (25.56 %). The host prediction was *Klebsiella* phage for VC250, and *Escherichia* for VC509. The viral cluster 375 was seen in UCb-Hv (53.69 %) but the host is unknown.

3.3.3 Network Analysis

The NetCoMi analysis determined the microbial association between each condition. In Figure 40, all associations are shown for healthy bacteria. The network included 29 bacterial genera and 104 viral clusters. A total of 1,807 links were drawn, which were separated into six clusters with bacterial associations and two clusters with only viral associations. The smallest cluster consisted of *Phalaenopsis* and two CRCE nodes. *Parabacteroides* were also associated with four common VCs. 11 common VCs were associated with *Longibaculum*, *Staphylococcus*, and *Acetatifactor.* In this cluster, two healthy nodes were also included. Two other healthy nodes were also present in a cluster containing *Prevotella*, *Yersinia,* and *Klebsiella*. That cluster had 29 viral associations, whereas one was from CRCA, four from CRCE, three from UC, 18 from common, and two belonged to healthy. The biggest cluster included those bacteria with the most connections (hubs): *Escherichia*, *Bacteroides,* and *Massilistercora*. It also included four hub VCs: VC243, VC39, VC445, and VC504. VC243 and VC445 had *Escherichia* as their predicted host, whereas VC 39 had *Citrobacter* and VC504 *Streptococcus*. In total, it had 20 bacteria and two nodes from CRCA, three nodes from CRCE, two nodes from Ulcerative Colitis, 24 from common, and two from healthy.

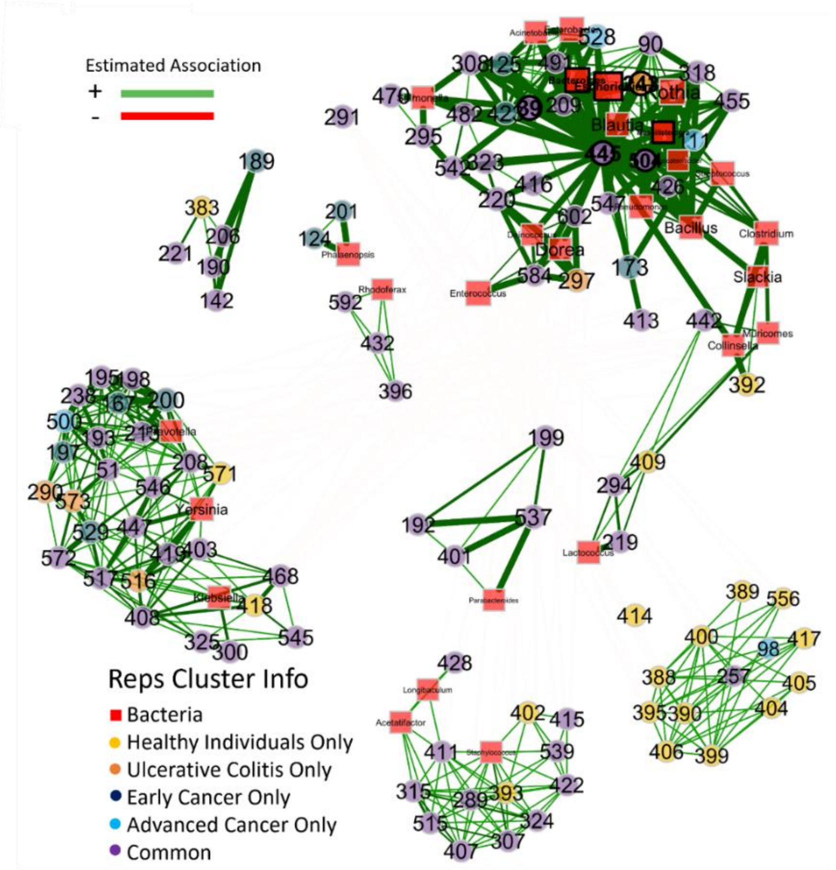


Figure 40_NetCoMi Network From Cross-Infections With Healthy Bacteria

Six clusters with bacterial associations were formed and two with only viral connections. The main hubs were *Bacteroides*, *Escherichia*, *Massilistercora*, VC39, VC243, VC445 and VC504. The predicted hosts were the following: VC39-*Citrobacter*, VC243 and VC445 – *Escherichia* and VC504-*Streptococcus*.

The second NetCoMi network was based on all connections made with bacteria from Ulcerative Colitis. The network split into two major clusters but was not fully disconnected. A total of 4,214 links were drawn between 55 bacteria and 136 viral clusters. In the left cluster, *Escherichia* was the hub connector, and in the right cluster, *Monoglobus* and *Massilistercora* were the main genera. The most viral connections (hubs) were seen for VC523, VC70, VC78, VC161, VC541, VC85 and VC458. VC161, VC85, VC458, and VC541 were classified as common viral clusters, whereas VC523 belonged to CRC early, and VC70 and VC78 were viral clusters from CRC advanced. All of the viral hubs were situated in the right cluster. The right cluster was generally more viral-heavy compared to the left cluster, which had more bacteria. Two hosts were predicted for all viral hubs: *Escherichia* (VC70, VC85, VC161 & VC523) and *Enterococcus* (VC78, VC458 & VC541).

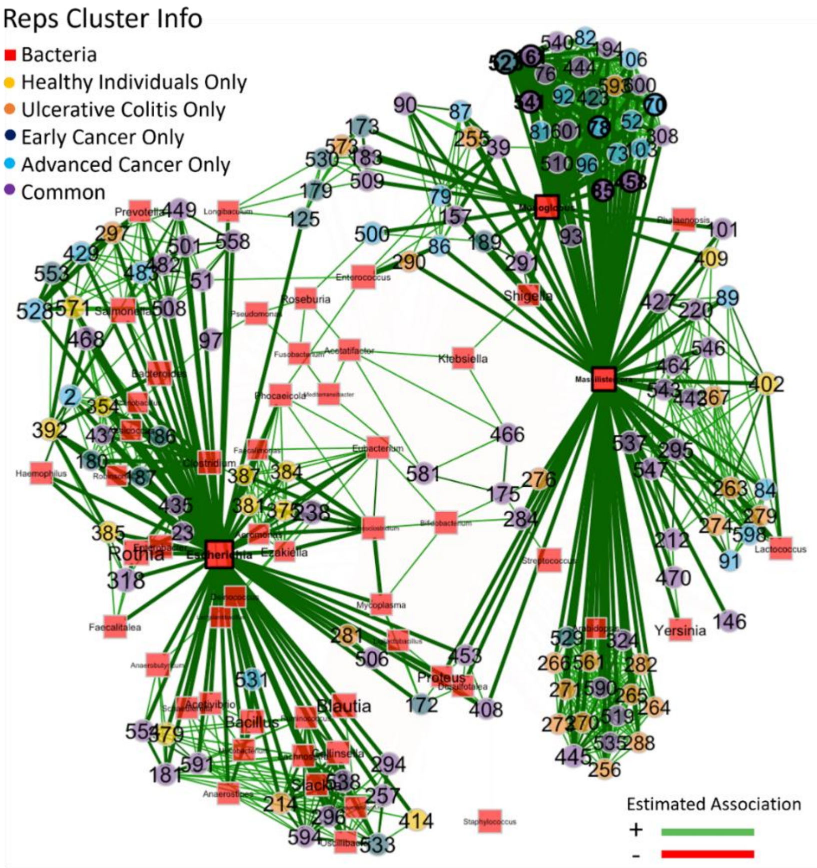


Figure 41_NetCoMi Network From Cross-Infections With Ulcerative Colitis Bacteria

The network had two clusters, but they were connected as well. Both clusters contained hubs, but the left one had only *Escherichia*, and the right one had *Monoglobus*, *Massilistercora*, VC70, VC78, VC85, VC161, VC458, VC523, and VC541.

A different network was presented when bacteria from early cancer were used as a base. Seven clusters were formed, whereas one had only viral connections. The others had 87 bacteria with 142 VC included. A total of 8,382 associations were detected between bacteria and viral clusters. One cluster included three bacteria, *Longibaculum*, *Neisseria*, and *Erwinia*, and eight CRCA VCs, seven CRCE VCs, 22 common VCs, and one Ulcerative Colitis VC. The second cluster was built around *Escherichia* and *Rothia* with three CRCE clusters, three UC clusters, one healthy, and seven common clusters. A small cluster was created with *Providencia* and one CRCA VC, two CRCE VCs, and three common VCs. *Ligilactobacillus* was connected to *Fusobacterium* and *Fusobacterium* to VC42 (common). A different cluster was made out of *Slackia*, *Plantactinospora*, *Erysipelatoclostridium*, and five Ulcerative Colitis VCs, one from healthy and two from common. The last cluster was the biggest cluster, which also included nine hub bacteria and three hub viral clusters.

The key bacteria were *Anaerostipes*, *Deinococcus*, *Desulfitobacterium*, *Eubacterium*, *Fusicatenibacter*, *Helicobacter*, *Lacrimispora*, *Roseburia*, and *Thermotoga* and the viral clusters were VC460, VC478, and VC425, which were detected in more than one condition (common). The predicted host for VC460 and VC425 was *Lachnospira* and for VC478 *Alistipes*.

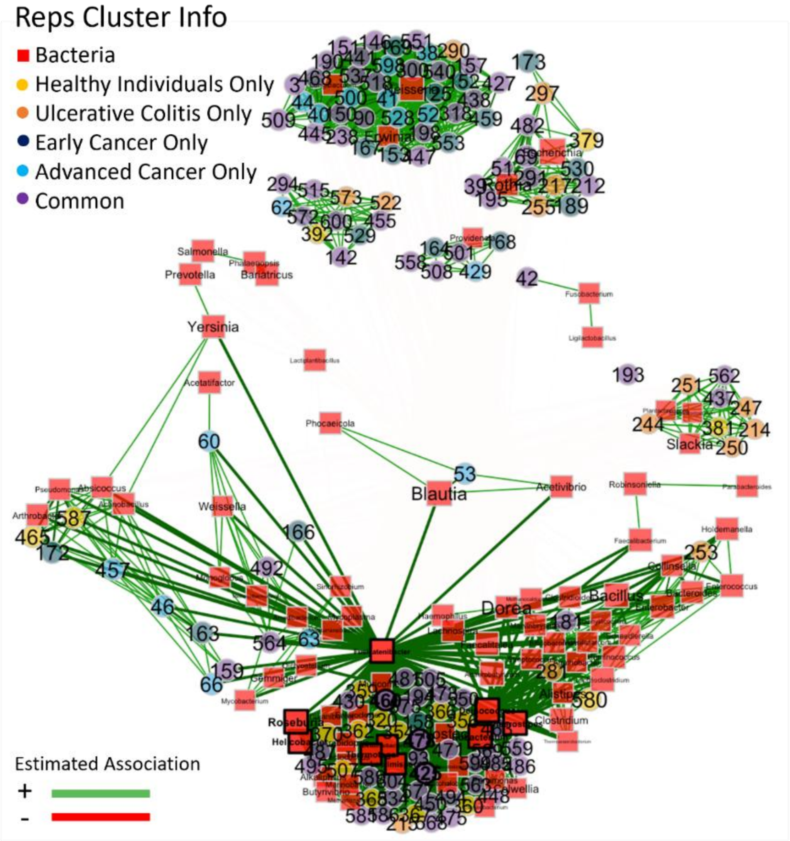


Figure 42_NetCoMi Network From Cross-Infections With CRC Early Bacteria

Seven clusters were created when CRC early bacteria were used. One cluster had only viral connections, and the other six did not have any key components included. The biggest cluster had all hubs combined: *Anaerostipes*, *Deinococcus*, *Desulfitobacterium*, *Eubacterium*, *Fusicatenibacter*, *Helicobacter*, *Lacrimispora*, *Roseburia* and *Thermotoga* and the viral hub clusters were VC460, VC478 and VC425.

The last network was based on advanced cancer bacteria samples. The entire network was connected with 6,212 associations and did not have a clear separation. It included 81 bacterial genera and 131 viral clusters. Three clusters could be determined. One cluster was connected through *Roseburia*, a hub bacterium. *Roseburia* connected *Calothrix*, *Acetivibrio*, *Faecalitalea*, *Collimonas*, *Synechocystis*, *Faecalimonas*, *Mageeibacillus*, and *Fermentimonas* with 14 CRCA viral clusters, whereas one (VC1) out of those 14 VCs was a hub. *Roseburia* also associated nine common viral clusters, which held two hubs (VC537 and VC590). One CRC viral cluster also belonged to the *Roseburia* cluster. The second cluster had *Escherichia* as the main connection point, which was also a hub. The last cluster had *Mediterraneibacter* as the hub, which also connected two other hub genera, *Bulleidia* and *Erysipelatoclostridium*, together with two hub viral clusters, VC144 and VC143. VC144 and VC143 belonged to CRC early. The host predictions for VC1 and VC143 came back as unknown. VC590 and VC537 had *Escherichia* predicted as their host, and VC144 had *Ligilactobacillus* as the prediction. The last identified hub was *Neisseria*, which is connected to *Escherichia*.

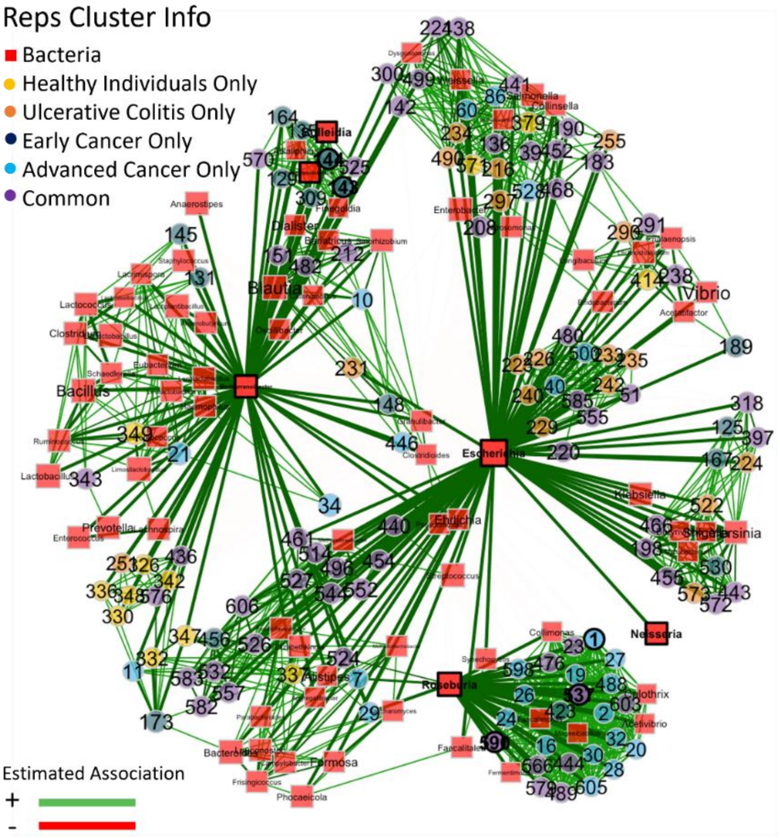


Figure 43_NetCoMi Network From Cross-Infections With CRC Advanced Bacteria

This network did not have a clear separation as everything was connected through something. Three different clusters could be determined, which were also the hubs: *Escherichia*, *Roseburia*, and *Mediterraneibacter*. *Bulleidia* and *Erysipelatoclostridium* were also identified as hubs but did not separate from the others. *Neisseria* was also identified as a hub. The viral hubs VC143 and VC144 were found close to *Bulleidia*. VC1, VC537, and VC590 were also hubs but were detected close to *Roseburia*. VC1 and VC143 had no host predicted, whereas VC590 and VC537 had *Escherichia* predicted, and *Ligilactobacillus* was the prediction for VC144.

4. Discussion

4.1 Targeted Single Phage Isolation

4.1.1 Developmental Challenges

Phage therapy is one of the alternatives to battle antibiotic-resistant pathogens. Phage therapy uses phages to treat patients with bacterial infections. Phages are viruses that are abundant worldwide [12]. However, phages are strictly host-sensitive, and finding the right phage for the patient's strain is a challenge despite the easy isolation process [14]. The current isolation process that scientists use is the old traditional plaque assay that Félix d'Hérelle developed. The process starts by mixing a bacterial culture with a viral source (wastewater, processed stool, etc) and incubating that overnight. Bacteria are separated via centrifugation from the supernatant, which might include virus-like particles. The supernatant is then tested via plaque assay for phages. Clear lysis zones in the assay indicate the presence of phages for that particular strain. Different plaque shapes suggest multiple different phages are in the supernatant. Each plaque shape needs to be purified before any characterization can start [33]. Traditionally, several rounds are performed of plaque picking – culturing – plaque assay – plaque picking – plaque assay, and so on until plaques are uniformed. This process can take days. In the subsequent step, DNA is extracted and sent for sequencing [116]. Up until the genome assessment is finished, all further analyses are on hold. If genome assessment determines that the isolated phage is pure and novel, scientists start with phage characterization; burst size, latency period, host range, morphology, and many more tests. The entire isolation and characterization process can take months, especially for complex pathogens. The process is long, laborious, and not well backed up with computational data [33]. There is currently not one phage reference bank, but multiple, and none of them include everything. Information is spread, computational tools are limited, and in their infancy. However, well-characterized phages are the basis for a successful and safe phage therapy treatment. If time is a critical factor, and the patient's life is on the line, tests can be done in parallel, but that saves only marginal time. At the bottom line, phage isolation and characterization take time and are a lot of work.

In my thesis, I developed a method that reduces the isolation process and initial characterization steps by hours. The method I used is called viral tagging. Viral tagging was initially created by Deng et al. [74] to investigate the phage-host interaction. Phage-host interactions and their consequences have been analyzed via viral tagging on multiple occasions, such as in marine environments or human stools. In the later, viral tagging was further extended to investigate phage-host relationships on a single-cell level. However, none of the previously mentioned studies have used viral tagging to isolate phages for phage therapy. As mentioned above, many phages are urgently needed for phage therapy and need to be found fast. I used the viral tagging protocol developed by Deng et al. [74], altered by Džunková et al. [115], and recreated a new one. The new method is now quicker than before as I reduced the time from 6-8 hours to 1-2 hours, depending on the complexity of the targeted bacteria, and created it in a way that it can be used on basically every bacterium. The method was tested for a wide range of pathogens, gram-negative as well as gram-positives, aerobic bacteria, and anaerobic as well as for human and mice stool bacteria in a consortium. Almost all the ESKAPEE pathogens were used: *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, *Escherichia coli*, and many more. Additionally to the variety of pathogens, the method was also created in a way that every generic flow cytometer can be used to separate viral-tagged cells from non-tagged cells. I tested the protocol on three different Beckman Coulter flow cytometers (MoFlo XPD, CytoFlex, and FC500), one from Sony and three from BD (FACS Canto, FACSAria, FACSMelody). Independent of the pathogen and flow cytometer, the method did well. The expected shift was seen in all experiments with slight variations from machine and pathogen, but a clear separation was present from the viral-tagged population compared to the non-tagged population. The successful establishment of this method enabled us to isolate numerous phages in a timely manner for a variety of pathogenic strains from different environmental sources, e.g., wastewater or patient samples. The amount of high-throughput isolation techniques is scarce. One reported technique is the HiTS (high-throughput screening) method. This method processes a sample size of >500 samples in parallel by using multi-deep well plates for phage amplification. A drop of each well is then spotted on a lawn to check for lysis. Despite its big sample size, the technique needs at least four days in a row to screen for lytic phages [117]. The targeted single phage isolation protocol isolates lytic phages within 24 hours and beats other traditional methods by time and sample size. Other high-throughput detections are metagenomic sequencing. Phages are detected, but they are not isolated in this process as the sample gets destroyed for DNA isolation. Additionally, a metagenomic sample contains multiple phages and not only one.

The singularity is ensured in the viral tagging protocol as single cells can be sorted via flow cytometers. However, flow cytometers have two major disadvantages. They are pressure-operated systems with a carrier fluid. The fluid carries the sample and aligns it for the lasers. By doing so, cells are put under a lot of stress and the carrier fluid holds a risk for cross-contamination. I transitioned to a microfluidics device to eliminate those external stress factors from my protocol. The device is pressure-free and a closed system, meaning the sample does not come into contact with any external fluid. The cell stays in its preparation buffer. An additional benefit from the microfluidics device is that it can be put into an anaerobic chamber enabling us to process the samples completely oxygen-free. Anaerobic sampling has been successfully done by Clavel's Lab [118]. When I compared the viral tagging data generated with a flow cytometer to the data generated by the microfluidic device, I saw little to no differences regarding the results. Apart from the graphical output, the size scatter plot showed the same cloud on both devices, and the fluorescence shift from viral-tagged cells was identified as well. The successful transition was confirmed when isolated tagged cells formed plaques in the subsequent plaque assay, and non-tagged cells did not.

After I successfully established the phage isolation protocol on both devices with a great time-saving aspect, we needed a method to quickly characterize the isolated phages to see if it was worth going forward with sequencing them or eliminating them from the characterization process. In phage therapy, only certain phages are desired. Phages need to be fast, have a high burst size, and need to be free from toxins or virulence factors [12]. The latency period and burst size are calculated via one-step growth curves. Toxins and virulence factors are determined after the sequencing process. I developed the targeted single phage isolation protocol to check for those desired properties. Viral-tagged cells are sorted individually in wells that were prefilled with the target bacteria. This step makes plaque purification rounds redundant and saves much time, as only one single phage is sorted per well, enabling us to screen at least 93 phages at once (96-well plate = 93 samples + 3 controls). Additionally, this step minimizes operator biases. Bacteria cells are pre-filled via the sorter into the wells. In the subsequent incubation step cell growth was monitored. Permanent monitoring allowed us to separate non-infective cells from active infection. This step eliminates phages, which might be temperate phages.

Phages that integrate into the bacterial genome and do not follow a lytic life cycle. It also highlights phages, which are slower in their infection as the point of killing is mirrored in the bacterial growth curve. For the proof-of-principle testing, I selected phages with a rapid active infection for the next characterization phase. However, as viral tagging enables us to sort thousands of phages at different time points, slower phages do not mean directly bad phages. The growth curves give us a first characteristic indication, but they mainly tell us if this phage infects the host. Wells with no phage infection can be omitted for the downstream procedure.

Continuing with the downstream process, the next step identified any unwanted traits, such as toxin genes or any other virulence factors. I established a multiplex qPCR with primers that amplify heat-labile toxin from *E. coli* and eae (intimin). Eae is a major virulence factor in Shiga-toxin-producing *E. coli*. It plays a major factor in bacterial adhesion on the intestinal wall [119]. Toxin genes, antimicrobial genes, or other virulence factors are often transferred by prophages to the pathogens via horizontal gene transfer, creating even more dangerous pathogens than before [76]. So, detecting those genes can be used as a landmark that the newly isolated phage might be a prophage, as their existence in bacteria have been found often in proximity to prophage-encoding regions [76]. Since the sequences of those genes are known, PCRs can be performed to detect those genes. qPCR has been used before to detect phages despite the absence of marker genes in phages. Either the targeted sequence was already known to quantify phage particles within the sample, or the targeted sequence was partially known [120], [121], [122]. I created primers based on the nucleic acid structure of conserved protein regions for the chosen toxin gene and virulence factor. Based on conserved capsid structural genes, I also designed primers to differentiate between the phage itself. Structural genes are well conserved as they are essential to protect the phage genomic information [123]. Compared to other qPCR results for phages, I did not quantify via qPCR. I was only interested in the presence of those genes. So, an amplification curve was enough for a positive test result. The qPCR test results for T1 and T7 were as expected, but for T4, one dilution series went wrong in the toxin curve. Error bars overlap with the entire bar. Despite that error, qPCR is a sensitive tool to quickly determine phage characteristics. Phages were identified at low concentrations and when mixed with other viral particles. These results align with other studies [122]. Sensitivity is an important property of laboratory techniques. Consequently, I challenged the entire procedure. A known phage concentration was added to a known concentration of virus-like particles isolated from wastewater. The targeted phage isolation protocol was executed as established. However, the T1, T4, and T7 recovery was only managed when the known phage was present a thousand times more than the other viral particles. These results did not align with what we detected during the qPCR process, as low concentrations were still amplified. We figured that the tested sample size with 93 wells might be too small to detect the T-phages at lower concentrations as wastewater has many other phages inside, which can also infect *Escherichia coli*. As T-phages were labeled with the same color as the other VLPs, there was no differentiation whilst sorting, but this should have happened during the qPCR. Consequently, to check if the small sample size was causing this low sensitivity in the process, I ran the exact same samples through the flow cytometer and accumulated one million cells. With one million cells, T1 was recovered in every tested concentration. So, we concluded that if multi-well plates are used, more plates need to be tested to bump up the sample size to ensure phage recovery.

A second reason why sample size mattered is phage individuality. During the analysis of the proof-of-principles runs from T1, T4, and T7, we found that there are four major growth patterns. The first pattern was a common bacterial growth curve, as expected from *E. coli*. There was no phage infection of any sort. This might be a sign of a rapid bacterial defense strategy. As phages and bacteria co-evolve over the years, they have established multiple different strategies and can recall them in a reoccurring infection [124], [125], [126]. The second pattern showed the phage's total destruction of all bacterial cells – no sign of bacterial growth at all. These curves can be explained by a phage infection that was already progressing or by their ability to be fast and efficient. T-phages, especially T1, are known for their efficiency [127], [128]. And Patterns 3 and 4 are hybrid states of the first ones. Pattern 3 had bacteria growing initially, reaching a peak, and then being killed by the phages. Pattern 4 had a small bacterial growth at the beginning, but then phages killed off almost everything immediately, but in the end, bacterial cells regrew. These hybrid forms could be explained by either bacterial defense systems that the phages overcame over time or different states of bacterial cells. Despite being processed together and handled completely the same, bacterial cell individuality also plays a role [129]. These four patterns have been seen in the tests with T1 and T4 whereas T1 experiments had more of pattern 2 and T4 more of 3 and 4. The difference can be explained by the phage infection ability as latency periods differ [130], [131]. To ensure that any protocol steps did not initiate those four patterns and were solely biologically based, cytotoxicity tests were performed. All three phages were put through the staining procedure, and titer was determined via double-layer plaque assay, and no decrease of the titer was seen. This aligns well with what the literature has reported [132]. Additionally to the toxic check, one-step growth curves were performed in 96-well plates in the shaking plate reader to mimic experimental conditions. One-step growth curves are traditionally done in one big flask, with one sample taken every minute. T1 was determined to have a latency period of around 13 minutes and a burst size between 60 to 100 progenies [133]. T4 was reported with a latency period between 12 and 25 minutes depending on the host strain and culture media with similar numbers for T7 [134], [135], [136], [137]. Despite the fact that latency period and burst size slightly vary depending on the host strain and culture media, the recorded data from the 96-well one-step-growth curves follows the literature. This rules out the idea that the protocol influences phage infection dynamics and leaves the assumption that phages in a community represent one view to the outside but have underlying heterogeneity where each phage acts on its own. Viral individuality is an important step to better understanding phages and making phage therapy safer [138].

4.1.2 Limitations of Targeted Single Phage Isolation

Despite the successful development of the targeted single phage isolation technique, which we believe will support phage isolation in various ways, the method does have its disadvantages. The biggest take is the equipment needed for the execution. A flow cytometer or microfluidics device is necessary to separate viral-tagged and non-tagged cells from the standard equipment such as incubators, lamina flow cabinets, or centrifuges. Next to the flow cytometer, a plate reader is essential with heating and shaking to ensure phage amplification and monitoring for the downstream process. Although a single-plate plate reader is sufficient, a multi-plate plate reader would boost the entire phage screening size. Since we have experienced that the monitored incubation is the bottleneck in this operation. Yet, those machines are not enough. For the last characterizing step, a qPCR machine is needed including costly qPCR kits. If we compare the developed method to the traditional double-layer plaque assay, the plaque assay is cheaper and easier. It is easy since the double-layer assay does not contain any advanced methods that need years to get experienced and well with, and cheap since it is mostly media and Petri dishes. In contrast, the targeted single phage isolation technique is costly and has the equipment that need experienced users to execute the protocols well. The operators also need to handle the maintenance with care and consistency. Additionally, a qualified understanding is essential to analyze the data and interpret the findings. Bacteria behave differently during the process, and strain variances must be differentiated from actual errors, which can be subtle. Those subtle shifts are especially important during the sorting process. In addition to the complexity of the procedure, we always have to remember that viral-tagged cells are not equal to an active infection. Same with the spot assay, not every spot contains active phages. It might be phages, but it also could be a lytic enzyme within the spotted solution. So, active infection can only be verified when single plaques are present in the plaque assay or bacterial growth is diminished during the monitoring. However, bacteria need to grow in agar as well as in multi-wells, which is tricky for some bacteria. Complex bacteria such as *Helicobacter pylori* will need protocol adjustments to enable successful growth monitoring. However, multi-well plates will be an excluding factor for some bacteria as limited media supply and gas exchange may not be guaranteed. A further limiting factor might be the primer availability. As phages do not have genetic markers, a more complicated approach has to be chosen at the moment to create those primers. However, multiple qPCR primers can be equipped with the same dye to save time and increase options to eliminate unwanted traits such as toxins. This would help distinguish unwanted phages from potentially helpful ones but would not identify the exact factor. However, as research advances and more and more facts are brought to light, a better and more suitable solution may come up.

4.2 Helicobacter pylori Phages and Their Diversity

Helicobacter pylori has infected over half of the world's population, with infection rates higher than 80 % in some areas. Higher infection rates were correlated with poor socio-economic standards, poor living standards, and lack of hygiene [139], [140]. Often, multi-generation households are more likely to acquire *Helicobacter pylori* as in-person transmission routes are supported. Oral-oral or fecal-oral transmission is highly likely as *Helicobacter pylori* DNA was detected in saliva, dental plaque, feces, gastric juice, and vomit [141]. Also, contaminated water has been reported as a source [142]. Once infected with *Helicobacter pylori*, the bacterium colonizes the stomach and remains there the entire life when left untreated. The infection stays asymptomatic for 90 % of infected people and never causes a problem. However, for the remaining 10 %, *Helicobacter pylori* infection can cause symptoms or become life-threatening. Symptoms such as gastritis are common. Nevertheless, 10-20 % of the patients can develop ulcers, whereas for 1-2 %, the infection can progress into cancer [143]. The reason is that *H. pylori* is spiral-shaped and equipped with flagella enabling it to screw itself through the mucus layer of the stomach [139], [141]. Behind that wall, the pH is much higher, and epithelial cells are directly accessible. Its virulence factor, VacA, forms pores through the epithelial wall, altering the immune system and cell apoptosis. The reaction cascade leads to chronic inflammation, resulting in lesions that can develop into cancer [139]. Cancer is the end stadium for many *H. pylori* patients, as gastric cancer was determined to be one of the deadliest cancer types globally [140]. Death can also occur through peptic ulcers. Ulcers are described as cracks in the mucosal wall greater than 3-5 mm [139], [140]. Deaths might be preventable with early diagnoses and the right treatment regimen. Diagnoses are made with urease activity tests, e.g., breath tests, or via a biopsy, which is used for histological analysis and culturing [141]. If positively diagnosed, there is mainly one treatment option: antibiotics. Current treatment regimens are two types of antibiotics and a proton pump inhibitor, and sometimes bismuth is added as a fourth [139]. However, as for every other bacterial infection, antibiotic resistance also skyrocketed in *Helicobacter pylori* strains. Clarithromycin, metronidazole, and levofloxacin have resistance levels up to 30 %. Alternative treatment options such as vaccines or others are currently under investigation. An alternative that has been shown to be successful for other pathogens is phage therapy [42]. Phage therapy uses bacteriophages, viruses, to eliminate the pathogen. Phages are the natural enemy of bacteria and occur at the same places as their hosts. Since phages are host-specific, only phages for *Helicobacter pylori* can be used in phage therapy. However, so far, only 57 sequences have been uploaded to the NCBI virus databank, and only four of them are complete and fully assembled [144].

In this thesis, we wanted to expand the amount of *Helicobacter pylori* phage sequences and understand them better. We have chosen *Helicobacter pylori* PMSS1 and SS1 strains for these experiments. As *Helicobacter pylori* is known for its high rate of genomic mutations throughout a year, DNA was extracted from the strains and sequenced, as laboratory culturing puts a lot of stress on the organisms [145]. The assembled sequences confirmed the identity of *Helicobacter pylori* PMSS1 and SS1. Their genomic length is around 1.6 Mbp with a GC-content of 39 %. This aligns with the strains deposited in the NCBI database [146], [147]. Also, the sequence similarity between them matches what has previously been described. One striking difference between PMSS1 and SS1 is that a big part of the sequence between 0.8 Mbp to 1.2 Mbp is inverted. The same inversion has been stated in J. L. Draper et al. [46] but this inversion is not permanent, same sequence region can be oriented as it is in PMSS1. After determining the correctness of the strains, viral-tagged cells were analyzed.

The current NCBI databank holds a total of 79 *Helicobacter pylori* phage entries, while the Bacterial and Viral Bioinformatics Resource Center (BVBRC) holds only 37 *Helicobacter pylori* phage entries [148]. In this thesis, one million viral-tagged cells were isolated for each strain, DNA was extracted and sequenced. A total of 22,331 contigs were assembled from the reads for both strains together. 11,429 contigs were assembled from the SS1 data and only 3,012 from the PMSS1 data. All of those contigs are longer than 1,000 basepairs as the shortest phage up-to-date has a sequence length of around 2,435 basepairs. We have used wastewater from different plants and stool samples from healthy individuals as viral sources. Overall, wastewater looks to be the better source than stool samples for isolation as more contigs came from that source. However, the difference is marginal. An explanation could be that in wastewater plants, thousands of stool samples are gathered and mixed, providing a bigger cohort than one single stool sample. As *Helicobacter pylori* is the most spread pathogen, its abundance in every plant is reasonable. So is the presence of *Helicobacter pylori* phages in the individual stool samples. An interesting observation was made when the amount of contigs isolated was compared based on which strain the isolation happened. More than 9,000 contigs were allocated to stool samples; roughly half were isolated with PMSS1 and the other half with SS1. A completely different picture was seen for wastewater. 2x more contigs were coming from SS1 samples than from PMSS1. The difference became more substantial after refining the analysis when only contigs longer than 10 kbp were considered. A total of 599 contigs were left for both strains, but none came from wastewater and PMSS1, but 251 contigs were isolated from wastewater with SS1. No significant allocation was determined between the strains for the distribution of contigs coming from stool samples. When we screened the NCBI database, no phage sequence had been deposited for *Helicobacter pylori* SS1 or PMSS1. Similar findings are in the BVBRC; no phages exist for the strains PMSS1 or SS1 [148]. The majority of deposited sequences were found from strains gained from patient biopsies in hospitals, for example, in Colombia or Portugal [44], [54], [149], [150]. Additionally, many of the deposited sequences were found in the genome of the *Helicobacter pylori* isolated strains as a prophage. Exceptions are KHP30 and KHP40. Those two phages were detected in the culture supernatant of the *Helicobacter pylori* strain. KHP30 and KHP40 were spontaneously freed [53]. Compared to this study, all viral-tagged contigs derive from wastewater or stool samples and not directly from the strain itself.

The assembly of viral-tagged contigs was difficult, as reference databases are scarce with information about *Helicobacter pylori* phages. To ease the process, an in-house database was created with every *Helicobacter pylori* phage sequence that was publicly available and every prophage prediction from every *Helicobacter pylori* strain. A phylogenetic tree was created based on Average Nucleotide Identity with *Helicobacter pylori* phages, prophages, the isolating host strains, Coronavirus, and isolated contigs from viral tagging. Coronavirus was the least related sequence compared to the rest, as it is a human pathogen and has no relationship whatsoever. A close relationship was seen between publicly available sequences and prophages predicted from publicly available sequences. As previously stated, the majority of deposited sequences were isolated from strains directly. So, deposited sequences are mostly prophage sequences. The close relationship aligns with the research that was conducted. A minor share of viral-tagged contigs were in close proximity to the publicly available sequences, but the majority had no relationship at all. Two small prophage groups separate from the public group ventures the guess that contigs close to those regions might also be prophages. Nevertheless, the majority remains unrelated and does not match any sequences available. When host strain information was added to the phylogenetic tree, no separation was visible between contigs isolated with PMSS1 and contigs isolated with SS1. No clustering was visible either when a heatmap was created based on ANI. This highlights the struggle for phage therapy, that even if sequences are closely related and show similarities, it does not mean that those phages might infect the host either. Phage sequences can be highly similar, but the hosts are not [24]. To further analyze the contigs, they were classified into three categories. Category 1 included all contigs where *Helicobacter pylori* was predicted as their host. Category 2 had all contigs, which had no host prediction or any correlation to any database (in-house or public). Category 3 held all contigs which had matches with sequences from a databank. A principle component analysis was conducted to see if phages within those categories are more similar to each other and if they can be associated with one strain. However, the analysis showed that the categories did not make a difference, and neither did the isolating host. These results underlie the general consensus that phages are highly diverse in their genome content and difficult to cluster. This goes along with previous studies that found clustering on the nucleotide level shows inconclusive results; however, on the protein level, clustering is possible [24]. Clusters were also seen when viral-tagged contigs were compared with publicly available sequences. Two major clusters were present on the PCA plot. Public sequences represented one group, and viral-tagged contigs formed the other group. For the public sequences, prophages and phages were tightly clustered together, whereas the prophages were more sparse than phages. Three prophages completely separate themselves from the remaining sequences, and a couple were found together with viral-tagged contigs. Five contigs separated from the group and did not belong to any of them. This PCA plot confirmed that viral-tagged contigs had little to no similarity on nucleotide level to currently available sequences. However, since a few prophages showed some overlaps, it raises the idea that these viral-tagged contigs really are potential candidates for phage therapy as no one has deposited those sequences before, to our knowledge. Especially four contigs seemed to have high potential. One contig came from a stool sample isolated with PMSS1; the other three were from Innsbruck wastewater and SS1. Their genome length was average. Gene predictions found phage genes belonging to the RNA and DNA metabolism and other enzymes essential for phage survival. However, the majority of phage genes were unknown. The reason behind that lack of information is the sheer greatness of genomic variation within phages for one gene, the constant evolution, and their mosaic structure. Additionally, phages do not have universal core genes, and only a limited amount of gene information was added to reference databases [151]. No gene clusters were found for 34 contigs from Category 2. A clinker visualization showed that only two contigs had one shared gene cluster. Both contigs were assembled from data generated by wastewater and the SS1 strain. This trend of dissimilarity persists throughout the study, which contrasts with what has been found for other phages. A comparative genomic analysis shows that phages that can infect *Escherichia coli* O177 share a high similarity of genome and proteome within themselves and other *Escherichia*-infecting phages [152]. However, comparing the *E. coli* study with our viral-tagged contigs, there is one major difference: the isolation process. In this thesis, *Helicobacter pylori* phages were isolated via viral tagging, sequenced as a community but then, divided during the assembly process. In the *Escherichia* study, *Escherichia coli* phages were first amplified individually and then also sequenced individually. Individual samples that are amplified allow for a deeper sequencing depth and also simplify the assembly process afterward. Sequencing depth is a critical factor in the assembly process, especially for metagenomic analyses. The more data is available, the deeper the subsequent analyses can go [153]. A second reason for the contrasting results might be the data availability in public databases. *Escherichia coli* as a bacterium and its phages are universally studied and highly studied. *Escherichia coli* is often used as a model organism to establish techniques and, therefore, often the first organisms with newly deposited data that was generated during that process. In the NCBI virus databank, over 5,000 nucleotide sequences have been placed for *Escherichia coli* phages, half of them being complete genomes and the other half being partially completed. In contrast, 79 sequences are shown for *Helicobacter pylori* phages in the NCBI virus databank. Out of those 79 sequences, only four are complete genomes. However, even if the amount of *Helicobacter pylori* phage data is poor, it still helps to analyze the unknown. We have used the incomplete data and screened it for auxiliary metabolic genes. AMGs are genes widespread in viral communities and associated with metabolic pathways, bacterial mobility, transportation, and biofilm production [154], [155]. Additionally to their involvement in bacterial cellular processes, AMGs are also relevant for bacterial virulence and can influence host health and disease [73]. In the thesis, we compared sequences from our in-house database with viral-tagged contigs. 16 AMGs were found in total, but only one was shared: NAD-dependent epimerase/dehydratase family. This family uses NAD as a cofactor to metabolize carbohydrates. Other enzymes that are relevant for carbohydrate metabolic pathways were identified in viral-tagged contigs, such as glycosyl transferase/hydrolase. Carbohydrates are a key element for phages during the infection process. Carbohydrates function as an entry port, but they also need to be eliminated during progeny release in the bacterial cell wall or are present in biofilm or mucus layers [156], [157], [158]. Auxiliary metabolic genes which were identified in publicly available sequences were more diverse. Genes were linked to DNA reactions, flagella and cell wall metabolism, and others. These can be explained as the currently available sequences are mostly from temperate phages that integrate into the host genome. Additionally, temperate phage sequences were also found in the bacterial DNA bioinformatically. Hence, the sequences were more complete than the viral-tagged contigs, enabling a better annotation.

Additionally to the AMG predictions, gene functions were also determined. In absolute numbers, viral-tagged contigs contain more gene clusters than what was found in our in-house databank. Especially more were found for DNA metabolism, AMGs, tail and others. For head, transcription, and lysis genes, the amount of gene clusters did not diverge that much. However, the picture changed when the same data was put in relation to each other. Many more genes were found for the in-house databank phages than viral-tagged contigs. Most genes belong to nucleotide metabolism, structural genes, and transport, which are all highly abundant in temperate phages. Additionally, since temperate phages integrate into their hosts, it has been discussed that they are the most sequenced entity [159]. Consequently, those sequences are available and can be used to annotate genes or compare and support assembly. In comparison, viral-tagged contigs are incomplete and sourced from wastewater/feces without any available details. Additionally, viral-tagged contigs came from uncultivated samples with low concentrations containing a mix of cells. All these facts led to smaller contigs. Smaller contigs are hard to annotate as many genes are bigger, such as structural genes, which range from five kilobase pairs to over 700 kilobase pairs [160]. This might be a reason why the overall amount of found genes and gene clusters is higher in viral-tagged samples compared to the in-house databank phages. These findings show that viral tagging exposes us to a completely diverse pool of phages we have never seen before. Viral tagging opens the possibility to study those phages that we usually lose during traditional phage isolation. The diversity of phages in viral-tagged contigs is also present in the unique gene comparison plot. The plot shows the comparison based on protein orthologous groups, so-called PHROGs. Viral-tagged samples separate clearly from the in-house databank samples. There is no overlap between those two groups, although the PHROG database contains 38,880 groups. Despite all efforts, the majority is classified as unknown and only associated with a number. Samples from the in-house databank were associated with PHROGs from head proteins, structural proteins, lysis, and transcription regulation. Nucleotide metabolism PHROGs were detected in both groups. Viral-tagged samples had additional PHROGs for tRNA methyltransferases and ATPases. Those differences can again be explained by the biases introduced between sample processing, sequencing, and assembly, as described previously. In addition to those biases, little is known about gene function in *Helicobacter pylori* phages, so it is possible that some of the detected phages are superior in metabolic functions, whereas others are better in propagation. If this is the case, the separation based on their properties can be explained. Another comparison based on PHROG was executed for endolysins. Endolysins are phage enzymes that can break down bacteria cell walls and are a promising antibacterial agent. Our in-house databank and all viral-tagged contigs were screened for endolysins, and 20 endolysins were found. However, only one was identified from the in-house databank. The remaining were all associated with viral-tagged contigs. Those 20 endolysins split up into 8 PHROG groups. Each group contains different amounts of protein sequences. The smallest group, phrog_15351, comprises 6 sequences, and the biggest group, phrog_7, has 3,145 sequences. For each phrog, suggestions are made for the closest protein family (PFAM). Depending on the phrog, some groups have rather vague protein families suggested, such as dipeptidase (phrog_2649), whereas others are more clear, such as phage lysozyme (phrog_7). Overall, all of the suggested protein families have activities to degrade the bacterial cell wall but slightly different techniques/target points, which are reasonable as all groups are classified as endolysins. Endolysins for *Helicobacter pylori* are currently only tested in an engineered form, as the pathogen has an outer membrane that protects the cell wall. Scientists created an enzyme called “Artilysin,” which is a two-component enzyme as scientists linked holin and endolysin together to break down the outer membrane [161]. The sequences for the enzymes were taken from KHP30 and 1961P phages. The artificial enzyme is expressed in *Escherichia coli* and has successfully been tested on *Helicobacter pylori*. Artilysins had a bacteriostatic effect [161]. So, expanding the database with endolysin sequences enables scientists to create more of those artificial enzymes and create an additional option to treat antibiotic-resistant *Helicobacter pylori* infection.

4.2.1 Current Limitation for *Helicobacter pylori* Phage Study

As with every technology, viral tagging also has its limitations. As previously described, a major drawback is the method's costs and its need for well-experienced operators. Despite the area of expertise and equipment, we could only identify the above-described phages, but we did not isolate them. We sorted one million viral-tagged cells and have used all of them for DNA extraction and sequencing to gain the most detailed picture. By doing so, we lost all of them and could not use any of them for phage amplification. So, in the future, a more sustainable process needs to be developed where phages can be identified and isolated quickly.

Furthermore, the bioinformatic process needs to be improved as many of our data came back labeled as unknown. We have created an in-house databank, hoping it would help with the analysis and shed more light on the dark matter as the publicly available data was missing, but most of the data remained unknown.

So, until phage therapy for *Helicobacter pylori* can be used, we must solve the isolation process and refine the bioinformatic analysis. The more phages we can isolate and identify, the more knowledge we can gain about *Helicobacter pylori* phages, which will help us in the future. Then, the potential of phage therapy is clearly given to revolutionize the treatment of antibiotic-resistant infections.

4.3 Colorectal Cancer Cross-Infection Study

4.3.1 Bacteriome and Virome

Over 29,870 research articles and 13,311 reviews about gut microbiome and disease have been published in the last few years [162]. The general takeaway from all of those is that there is a link between the human gut, its colonizers, and the host's health. The exact mechanisms behind the interplay between host and microbiota are often unknown and are still the circle of attention. The attention is on better understanding the interaction between bacteria and disease to develop therapies or support early diagnostics. Despite major efforts and new technologies, the gut and its disease remain a mystery that is waiting to be uncovered [162].

In this thesis, we used the developed targeted single phage isolation technique to determine which phages are present in certain conditions and how those phages can influence their surroundings. We had three healthy stool samples, three from patients with Ulcerative Colitis, three from patients with colorectal cancer in the early stage, and two from cancer patients in an advanced stage. First, we wanted to know each condition's bacterial and viral composition. Whole genome sequencing was conducted, and bacteriome and virome were determined. The bacterial analysis showed a clear picture between healthy individuals and disease. The healthy samples were dominated by *Escherichia* spp, followed by *Bacteroides*. *Escherichia* is a gram-negative, rhode-shaped bacterium that has been detected in over 90 % of the world's population [163]. It belongs to those bacteria that colonize the gut at first and are widely present. *Escherichia* are facultative anaerobes, meaning they can survive in oxygen-free and aerobic surroundings. *Escherichia* eliminates the remaining oxygen in the gut and provides a complete anaerobic system for strict anaerobes [163]. Due to its supportive actions, *Escherichia* are categorized as commensal bacteria. The majority do not harm their hosts but rather support it. Exceptions are, for example, enterotoxigenic *Escherichia coli* or enterohaemorrhagic *Escherichia coli*, which can be deadly for the host [164]. Many different types of *Escherichia* are collected mainly by food, environments, or animals during a lifetime, explaining its dominance in healthy individuals' bacteriomes. In a healthy gut, the microbial composition is balanced and in a mutual relationship with the host. If the balance gets out of control and the microbial gut composition changes, it becomes dysbiotic, and the host's health suffers [55]. These microbial changes can be seen in the bacteriome analyses of Ulcerative Colitis, CRC early, and CRC advanced. The microbial composition differs highly from the individual one. *Escherichia* is no longer the dominant genera; *Bacteroides* has taken over. The lack of *Escherichia* in Ulcerative Colitis samples aligns with scientific reports in the literature, where no association between *Escherichia* and Ulcerative Colitis has been seen so far [165]. With regards to the cancer samples, it is reported that cancer patients presented higher *Escherichia* abundance than their controls [165], [166]. *Escherichia* is the second most prevalent genus in our cancer samples, but it is less abundant than in our healthy controls. Despite the differences in health control, it is clear that *Escherichia* plays a more crucial role in CRC than it does in UC, as the levels are more elevated in CRC. Literature states that in certain strains of *Escherichia coli*, genes that destroy the DNA are present. Those specific strains have been increasingly found in tissues around tumors in the GI tract [166]. Also, increased levels of *Bacteroides* were seen in all three disease conditions. *Bacteroides* belong to the normal gut microbiota and exist in at least 20 different species. It is an anaerobic and gram-negative bacterium that has toxic and non-toxic types. Toxic *Bacteroides* have the Bft gene that produces fragilysin. Fragilysin damages tight junctions and destroys mucosal barriers, causing inflammation [167], [168], [169]. The increased abundance of *Bacteroides* in all three disease conditions aligns with the literature. Its presence in the healthy samples also makes sense, as *Bacteroides* is part of the normal gut flora [167]. Another member of the gut microbiota is *Clostridium*. *Clostridium* is gram-positive and anaerobic. Despite being part of the normal gut microbiota, it has some species that can cause dysbiosis, such as *Clostridium difficile* [170], [171]. *Clostridium difficile* is one of the leading causes of Ulcerative Colitis and Crohn's disease. This fact is reflected in the data, as *Clostridium* is present in the Ulcerative Colitis and colorectal cancer samples. It has been reported that *Clostridium* was also found in CRC patients [172]. A different pathogen that is associated with CRC is *Pseudoflavonifractor*. Studies have shown that *Pseudoflavonifractor* is elevated in CRC patients compared to healthy samples [173]. The same picture can be seen in our data where the healthy sample has a small portion of *Pseudoflavonifractor*, whereas CRC early has already higher levels and CRC advanced has the highest abundance. Many aspects that other scientists have reported about the gut microbial composition in CRC and Ulcerative Colitis have also been reflected in our data. That is an outcome worth mentioning since our data only comprises three samples for each condition, respectively two for CRC advanced. Despite the small sample size, each condition is distinguishable. It also gives the impression that all three diseases are related as their bacterial composition is similar, except that some genera increase or decrease with disease progression. This impression can be strengthened by literature reports that confirm that Ulcerative Colitis can progress into colorectal cancer [174]. Additionally, first efforts have already been made to decode the microbial composition in diseases so that the microbial signature can be used as biomarkers during the diagnostic process [175]. One fact that is essential not to miss when determining biomarkers is to investigate every aspect of it. Early gut microbiome research mainly addressed bacterial composition. However, the youngest results have shown that not only bacteria but also viruses are responsible. The viral composition is as important and plays a crucial role in disease development. Hence, we also looked into the virome of each condition. However, due to the new taxonomic classification system, the majority were classified as *Caudoviricetes* and *Malgrandaviricetes*. Consequently, we created an UpSet plot based on vOTU viral clusters. The majority of viral clusters were unique for each condition, and only a few were shared with each other. Shared clusters were found in healthy and disease conditions but also between disease conditions or all of them; no mix stood out. That kind of diversity is expected in phage research as phage genomes are highly diverse and lack common sequences. Additionally, due to their small genome size, many sequences are fragmented and incomplete, which creates a challenge for comparisons. With continued efforts and the development of new bioinformatic tools, gaps will be closed eventually [24].

4.3.2 Cross-Infection and Network Analysis

The role of bacteria is well-studied in different disease settings, but the role of bacteriophages, which are more or as abundant as bacteria, is not well-studied at all. Therefore, we isolated bacteria and VLPs from healthy individuals, Ulcerative Colitis patients, and colorectal cancer patients in early and advanced stages and cross-infected them with each other. This allowed us to determine the bacterial and viral community in each condition (healthy and disease) and investigate the role of phages in each disease and how phages interact. Starting with the bacterial community in control groups (bacteria that got infected with their own VLPs), we see that the majority of bacteria are shared between all four conditions. The group consists of *Escherichia*, *Bacteroides*, *Enterobacter*, *Enterococcus*, *Clostridium*, and *Blautia*, which are all part of the normal intestinal microbiota and are considered commensal. The presence of those genera is expected in all conditions [163], [167], [170], [176], [177], [178]. Interestingly, only the control group of CRCAb-CRCAv has bacteria that are not shared with the others. *Lentilactobacillus* and *Granulibacter* were only detected in the CRCA control group. *Lentilactobacillus* are gram-positive, fermentative, and facultative-anaerobic. They belong to the lactic acid bacteria group and are considered probiotics [179]. *Lactobacilli* have been reported to help against cancer as some strains can modify the GI wall, support immune responses, start apoptosis, and metabolize molecules with antiproliferative and anti-inflammatory traits [180]. As little is known about the patient and their treatment at the hospital, it might be that the patient received probiotics to manage some of their symptoms. With regards to *Granulibacter*, *Granulibacter* belongs to *Acetobacteraceae* and is gram-negative. It can be found in the environment and is non-pathogenic to humans. Latest research, however, has shown that some strains can become problematic, especially for people with immune system issues or intravenous accesses [181]. Cancer patients qualify for those categories as chemotherapy or radiotherapy weakens the immune system, and venous accesses are needed [182]. As those two bacteria were only detected in CRC advanced, it can be hypothesized that those patients have undergone multiple rounds of treatment already, whereas the patients from CRC early just started. The bacterial composition from CRC early also indicates that changes are happening as only core bacteria were detected together with *Rothia* and *Neisseria*. *Rothia* is shared with the healthy sample and *Neisseria* with the CRC advanced. So, it could be hypothesized that changes are occurring. For the Ulcerative Colitis sample, it shows shared bacteria with the healthy control group but also with the cancer control group. Since Ulcerative Colitis is a relapsing and remitting disease, symptoms are not always present; they come and go. It was also reported that the bacterial gut composition changes during flares and remission. Patients in long remission present microbial communities closer to healthy individuals than flare patients. Diversity and richness change based on disease stage [183]. When we compared relative abundance in all cross-infection samples, we saw that *Escherichia* is the dominant bacteria in the control group but also in all the other cross-infections. Whereas the abundance in all cross-infections with CRC early bacteria have less *Escherichia* present than the others, especially CRCAb-UCv, UCb-CRCEv, and UCb-Hv. This contrasts with what is stated in the literature, where more *Escherichia* was detected in CRC than in healthy [165]. Colorectal cancer has also now been reported to have an association with *Neisseria* and oral microbiota [184]. The

Neisseria was more abundant in our dataset as well. Another present bacteria in CRCAb-Hv is *Bacteroides*. It is more abundant in that sample than in any other sample. *Bacteroides* has been linked to colorectal cancer, especially *Bacteroides fragilis*, which has the mucosal wall damaging toxin [167]. *Enterobacter* was also seen elevated but in samples from CRC in the early stages. Research suggests that *Enterobacter* is associated with cancer development, especially some strains that can start tumor growth and cancer progression [185]. Since *Enterobacter* was only found to be more abundant in early-stage CRC, it would fit the description from the literature. A different bacteria that was only more abundant in UCb-UCv is *Yersinia*. *Yersinia* is a gram-negative bacterium, has 11 species, and is classified as *Enterobacteriaceae*. Three of the species are harmful to humans and can, for example, cause gastroenteritis. *Yersinia* infections often look the same as a UC flare or can cause lesions, which are usually seen in Crohn's disease [186], [187]. Differentiation is hard as cases have also been reported that patients have Ulcerative Colitis and a *Yersinia* infection [188]. Overall, all mentioned bacteria have been reported with an association to some extent with one of the diseases tested and align with the literature. A contrary picture was seen for viral determination. A less distinct picture was drawn when viral clusters were determined and compared in self-infection and cross-infection groups. Within the self-infection group only one viral cluster was shared that was VC238, which was predicted to be *Escherichia* phage. Since *Escherichia* was identified as the most dominant genera, an *Escherichia* phage is expected. This finding aligns with the expectations and common knowledge about *Escherichia* and its phages [189]. Phages are unique and highly diverse. This phage diversity can also be seen in our data, as many viral clusters are not shared. CRCAb-CRCAv has the most amount of unique VCs. Half of the viral clusters have their hosts identified as unknown, the other half got *Lactobacillus* and *Limosilactobacillus* as their predicted hosts. *Limosilactobacillus* was also determined as one of the most abundant taxa in the CRCAb-CRCAv cross-infection sample which matches the viral data. Phages are only present when a suitable host is available, which is confirmed in the data. CRCEb-CRCEv had two unique viral clusters VC52 and VC157. Both viral clusters had *Salmonella* predicted as their hosts. *Salmonella* was recently identified as a driver of colorectal cancer. *Salmonella* interferes with signaling pathways, which are essential for the acetylation of proteins. Wrong acetylation supports tumor invasion and leads to a poor CRC prognosis [190]. Scientists also found that CRC patients have an increased level of

Salmonella phages [72]. That increased level of *Salmonella* phages in CRC patients is also present in our data and aligns with the literature. Viral clusters that were shared between all three diseases had *Isoptericola* and *Escherichia* predicted as their hosts. *Isoptericola* is a relatively young genus, with only 11 species included, which little is known about. On the contrary, *Escherichia* is very long known, well studied, and abundantly present. *Escherichia* phages have been reported to be more abundant in Ulcerative Colitis tissue than in healthy tissue and are suspect of investigation [191]. In colorectal cancer, scientists are still debating how *Escherichia* phages influence the disease. Phage increases as well as decreases of *Escherichia* phages have both been seen [192], [193]. These contrary findings are also mirrored in our data as VC173, which is predicted as *Escherichia* phage, has its highest abundance in Hb-CRCAv, followed by UCb-CRCAv and Hb-CRCEv but is not abundant at all in the cancer control groups or cross-infection samples which had bacteria from cancer patients involved. These results suggest that *Escherichia* phages are influential. To see which gut members are influential, we have used NetCoMi.

NetCoMi is the abbreviation for Network Construction and comparison for Microbiome data. NetCoMi visualizes associations between different microbial communities or samples [194]. In our case, NetCoMi created networks for healthy bacteria and all cross-infections, which were done on healthy bacteria, the network for Ulcerative Colitis bacteria and all cross-infections, the network for CRC early bacteria with all cross-infections and for CRC advanced bacteria with all cross-infections. Overall, the generated networks for all four conditions differ visually significantly. Also, the hub taxa and hub viral clusters (hub = point with the most connections) show little to no overlap. The evaluated hub taxa revealed that *Escherichia* is one of the hubs in the healthy, Ulcerative Colitis, and CRC advanced network. *Massilistercora* was determined as a hub in the healthy and Ulcerative Colitis samples, whereas Roseburia was identified as a hub in both cancer samples. All other hub taxa were unique for the condition, and so were all viral hubs. The network for healthy bacteria contained multiple separated clusters. The biggest cluster included all hub taxa: *Bacteroides*, *Escherichia, and Massilistercora*. *Massilistercora* was first detected in 2020 and described as being part of the phylum *Bacillota*. The new bacterium was isolated from stool, but little more is known about it [195]. *Bacteroides* and *Escherichia* are part of the normal gut flora and are widely present, which explains their many connections within the network. A healthy gut network only works if every communication and signaling cascade is kept upright and going. Network communication only works best if everyone is included and no one stays isolated. For that reason, the hub viral clusters align perfectly as the majority belong to the common group and are not skewed into one condition. Also, their host predictions fit, which were *Escherichia*, *Streptococcus*, and *Citrobacter*. All represented in the gut. The only exception is the hub VC which belongs to Ulcerative Colitis. However, Ulcerative Colitis patients in remission reported having similar gut microbiota as healthy individuals [183]. And since we do not know anything about the patient behind our samples, it could be a patient in remission. This hypothesis would go nicely as the interaction network from Ulcerative Colitis bacteria looks completely different compared to the healthy network. The Ulcerative Colitis network is connected throughout and has no separated clusters. Every member of the network is somehow connected with each other. The disrupted gut microbiota can explain this cross-talk. In the event of a disruption, the communication could be obstructed between the symbiotic bacteria, sparking a signaling cascade that leads to negative metabolic reactions or harmful immune responses, as inflammation can be triggered. Additionally, due to dysbiosis, the bacterial concentration alters pathogens' growth [196]. The different composition of gut microbiota in Ulcerative Colitis explains the different connections compared to the other networks. Next to *Escherichia* as a hub, *Monoglobus* and *Massilistercora* were also identified as hubs. *Monoglobus* is also found in the gut and is known for its pectin-degrading properties [197]. Pectin has been reported to benefit IBD patients as it can be metabolized into short-chain fatty acids, which are known to reduce inflammation and support gut balance [198]. These beneficial traits of *Monoglobus* could be the reason for an increased interaction between the bacterium and the remaining gut microbiota. In contrast, *Massilistercora* is not highly studied, so few connections can be drawn. However, since the hubs from Ulcerative Colitis are similar to the healthy ones, we could hypothesize that the patients might be in remission or on the way to remission. Studies have shown that patients in long remission are more likely to have a balanced microbiota than those with a flare [183]. Also, the viral hub analysis strengthens this idea as the host predictions were either *Escherichia* or *Enterococcus*. Both are common members of the gut, and their phages regulate their population and try to maintain a healthy balance.

This idea is also backed by the fact that in both cancer conditions determined hubs are completely different, as their gut changes are irreversible. For the CRC early bacteria network, the clustering is again more distinct like it was in the healthy network, but the determined bacterial hubs are unique. The bacterial hubs were *Anaerostipes*, *Deinococcus*, *Desulfitobacterium*, *Eubacterium*, *Fusicatenibacter*, *Helicobacter*, *Lacrimispora*, *Roseburia*, and *Thermotoga*. *Anaerostipes* are known gut bacteria, which are gram-positive and anaerobe. Their main activity is to convert carbohydrates into short-chain fatty acids. It can be hypothesized that the communication to *Anaerostipes* is increased to boost short-chain fatty acid production to reduce inflammation. However, studies have shown that *Anaerostipes* abundance is decreased in CRC and that some *Anaerostipes* species are more harmful to the host than beneficial [173], [199]. Another less beneficial member of the gut is *Desulfitobacterium* [200]. *Desulfitobacterium* resides in the anaerobic areas of the gut and takes H2 as an electron donor [200]. Microbial sulfidogenesis has been determined to be damaging as H_2S causes inflammation, stops apoptosis, and, as a result, leads to CRC [201], [202]. CRC is also supported by *Helicobacter*. *Helicobacter* colonizes the stomach and not the colon, but studies have shown that *Helicobacter* promotes CRC, and CagA is also associated with CRC [203], [204], [205]. A CRC-promoting association was found with *Eubacterium*. Studies have shown that *Eubacterium rectale* produces a toxin that triggers the reaction of the NF-κB (nuclear factor kappa B), which leads to inflammation and can result in cancer. However, there are other *Eubacterium* species such as *Eubacterium callanderi,* with the total opposite effect. *Eubacterium callanderi* supports apoptosis and cell-cycle arrest and is proposed to be used in therapy. Also suggested for therapy is *Fusicatenibacter* since *Fusicatenibacter* was less abundant in CRC patients than in healthy individuals [206]. A bacteria that were also decreased in CRC patients was *Roseburia*. *Roseburia* was detected with altered levels in IBD and CRC patients. It is an anaerobic, gram-positive, and rod-shaped bacterium that produces butyrate. Butyrate has been seen to reduce inflammation and keep the energy balance by maintaining levels of immune cells and cytokines. It was also reported to suppress tumor growth [207], [208]. *Roseburia* was also determined as a hub in the cancer advanced network. Despite being a hub in CRC early and CRC advanced, it looks like *Roseburia* interacts with a more distinct composition of microbiota in CRC advanced than in CRC early. In CRC early, *Roseburia* was part of the biggest cluster, which included all bacterial and viral hubs, and in CRC advanced, it was rather separated. This separation could be explained by the fact that the gut microbiota has been disrupted for longer in CRC advanced than in CRC early, and therefore, certain connections have been formed and stabilized. In CRC early gut microbiota has started to disrupt and interactions are happening between many members and are ongoing. So the difference between CRC early and advanced could be explained by time. This idea can be strengthened by the fact that the CRC advanced hubs are not similar to those determined in CRC early. The CRC advanced hubs are *Escherichia*, M*editerraneibacter*, *Neisseria*, *Bulleidia,* and *Erysipelatoclostridium. Escherichia* is present in healthy individuals as well as in patients. However, as long as the gut microbiota is balanced, the levels of *Escherichia* are optimal, but as soon as the gut microbiota gets disrupted and competition decreases, *Escherichia* can thrive. Due to the fact that it can proliferate in oxygen-free as well as oxygen-present environments it can take over many areas inside the gut [209]. Also found in the gut is the *Erysipelatoclostridium. Erysipelatoclostridium* has been reported in studies to be more abundant in CRC and IBD patients [210], [211], [212]. Also related to IBD was the third hub, *Mediterraneibacter*. *Mediterraneibacter* is a gram-positive, non-motile bacterium that can only live in obligate anaerobic environments. Its species *M. gnavus* was found in approximately 90 % of the tested human fecal samples and was associated with IBD, irritable bowel syndrome, and heart disease but not directly with CRC [213]. Also not directly associated with CRC was *Bulleidia,* another hub. *Bulleidia* is found in the gut, but mostly, it is identified with oral cavity infections rather than gut diseases [214]. These less common and obvious connections can be explained as viral tagging reveals interactions between bacteria and phage and is not based on abundance. Another example is *Neisseria*. *Neisseria* was identified as a hub in the CRC advanced network but is usually found in the upper respiratory tract or genitals as *Neisseria gonorrhea* causes the sexually transmitted disease gonorrhea but might have some connection to CRC [184], [215], [216]. Other examples were *Deinococcus*, *Thermotoga*, and *Lacrimispora* in the CRC early network, which have little to no connection to the gut and are less studied compared to others, such as *Escherichia*. Studies are also missing for phages and their connections inside the gut. We determined *Lachnospira* as a viral host in the early cancer network. A high abundance of *Lachnospiraceae* was found to be related to a high Immunoscore in CRC advanced patients. The Immunoscore is used as a biomarker for CRC prognosis [217]. A higher bacterial abundance would suggest a higher phage concentration; however, none of the predicted hosts for the CRC advanced viral hubs have *Lachnospira* as the predicted host. They had *Ligilactobacillus* and *Escherichia,* which are not inherently CRC-specific. However, since early CRC can progress into advanced cancer, phages from early stages could be used as a prognostic marker for disease progression. Nonetheless, robust biomarkers have not been determined yet.

4.3.3 CRC Cross-Infection Study Limitation

In this study, we delivered preliminary and descriptive insights into phage-host interactions in gut diseases. The reason for that is the unconventional method. We used unknown communities in an anaerobic viral tagging setting, gathered 100 cells, and sent those without DNA extraction. Due to this low and unorthodox input, the sequencing depth was affected and led to a short viral contigs recovery. Those short contigs challenged the assembly process. To address this challenge, the viral contigs were mapped to the gut phage database (GPD database). We could have avoided those issues by sorting each viral-tagged cell separately and amplifying them separately. Single-cells would also account for different bacterial concentrations. Viral tagging does not differentiate between different bacteria. If the sample holds an over-proportional concentration of one bacterial species, viral tagging will sort them as long as phages are available that can infect that species. However, single sorts carry other difficulties, such as contamination and financial burden.

Another challenge that we faced during the analysis of the CRC project was the availability of bioinformatics tools. Many bioinformatic tools are not built for such high-dimensional data. This has led to some irritable data, such as the *Neisseria* hub in the CRC advanced network. *Neisseria* was determined as a hub based on the calculations from NetCoMi, but due to the thresholds applied, *Neisseria* showed up on the network with only one connection to *Escherichia* despite being a hub. In addition, many bioinformatic tools are made to process a large group of samples, which enables statistical testing. Since our self- and cross-infection samples were only available once, the technique also reached its limits there.

5. Conclusion and Outlook

Overall, we can conclude that the developed technique showed great potential. The method enabled us to study phages on a single-cell level, examine their proteome under certain conditions, and look at their transcriptome at various time points. The discovery of heterogeneous infection dynamics of phages has shown that we still miss crucial information about phage behavior. The fact that phages behave differently on a single-cell level than they do in a community tells us that there are so many more underlying mechanisms going on that we still do not know about. Viral individuality must be further examined to fully grasp their influence on their environment. The developed technique allows us to do exactly that in a cultivation-free and unbiased way. We can only build stable, successful, and reproducible phage therapy trials that can be transferred to human patients if we figure out how this cascade of phage reactions works. For that, many more experiments must be conducted, starting with more bacteria. Currently, the technique was only performed on *Escherichia* and its T phages. The technique must now be tested on other pathogens. I would expand the technology to those bacteria that already have phages available with substantial background information. Potential candidates could be *Pseudomonas* or *Acinetobacter*. Those pathogens have phages available and are well-cultured in laboratory settings. If those are successfully established, the technique can be expanded to more demanding pathogens, such as *Helicobacter*. For *Helicobacter*, the amount of available phages is still limited, which could pose a challenge. Additionally, more sophisticated equipment must be purchased as *Helicobacter* only grows in microaerophilic atmospheres. However, the limited phage availability for *Helicobacter* could be tackled by expanding the knowledge gained from the second project.

In the second project, we uncovered some hidden knowledge about *Helicobacter pylori* phages. We used viral tagging to sort millions of cells, which were then bioinformatically analyzed. The data showed a completely different pool of phages compared to the currently available phage sequences in the public database. That data confirmed that the viral dark matter is still gigantic. The data also showed that despite the genetic similarity of the *Helicobacter pylori* strains, the collected phages are highly diverse. So, we concluded that the best possible scenario for *Helicobacter pylori* phage therapy is to isolate as many phages as possible with as many different strains as possible. We have to isolate more to achieve a sufficient and deep knowledge of *Helicobacter pylori* phages. Here, we would perform viral tagging experiments with many more *Helicobacter pylori* strains and wastewater as we determined it to be the best source. However, instead of sending every sorted cell for sequencing, some could be used for phage culturing. This way, we can identify phages computationally but should also have them physically in culture. With this procedure, we believe that the amount of phages for *Helicobacter pylori* can be drastically increased, and potential candidates for phage therapy can be chosen. In addition, phage research on *Helicobacter pylori* would also boost our knowledge about *Helicobacter pylori* and its diseases.

Since little is known about phages and diseases, viral tagging can also be used to support this research. Viral tagging helped to examine the interaction between phages and their hosts in disease conditions. In our third project, we were able to show preliminary connections. Many connections are still unknown as the phages behind them are still unknown. However, the data aligns well with existing literature. This fact supports our conclusion that the developed process is a great foundation for more and that phages play a major role in diseases and disease onsets. We endorse the idea that phages can function as biomarkers and contribute to prognosis. To substantiate our preliminary CRC results, I would suggest expanding the sample size. First, we need to collect more patient samples, maybe even from different cohort centers, and the according metadata to provide a solid base. Then, I would either do single-cell sorting with deep sequencing if finances are not an issue or sort more cells and do high-resolution metagenomics. Either way, sample size and sequencing quality need to be improved. A second strategy could be longitudinal studies. If patients are followed over a specific period of time, microbial compositions, abundances, and interaction networks could be monitored and mapped to the disease stages. With longitudinal data, we should be able to distinguish flares, remissions, and cancer progression and might predict them, which could lead to a different health prognosis. Additionally, the knowledge of phage-host interactions on a single-cell level is important as treatment outcomes can depend on it. Fecal microbiota/viral transplantations can become more targeted as the composition of the transplant can be optimized for the patient's microbiome. Allogeneic hematopoietic stem cell transplantation could also become more predictable when microbial dynamics are better understood.

5.1 Essence

We believe that with all these generated insights -on single phage characterization, phages for *Helicobacter pylori,* and phage interaction networks in Ulcerative Colitis and CRC-we have built a solid foundation. Further research options are endless, but all of them are equally important as they contribute to the greater good – to save people's lives.

6. Appendix

6.1 Supplementary Information

6.1.1 Targeted Single Phage Isolation

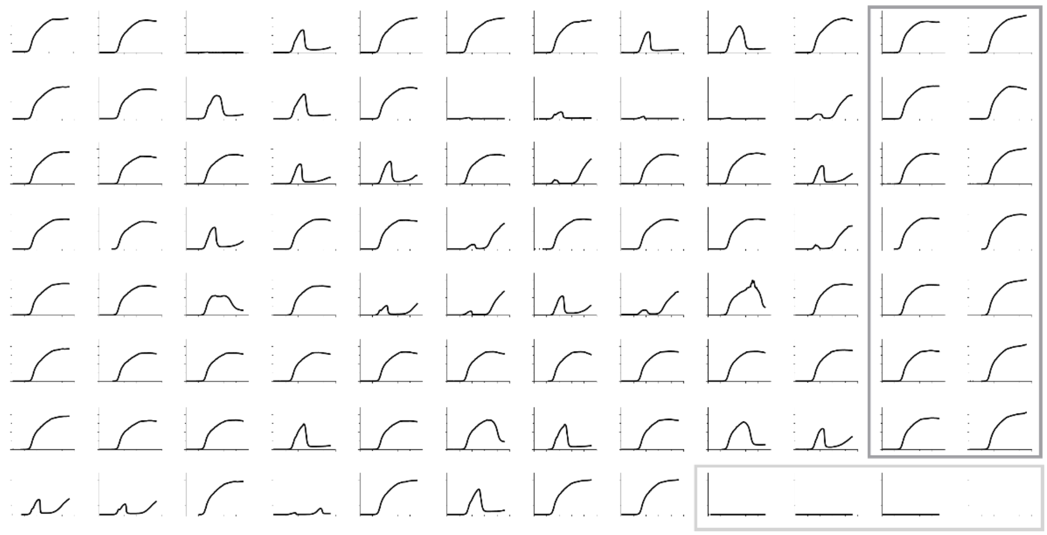
Supplementary Table 1_Primers for qPCR

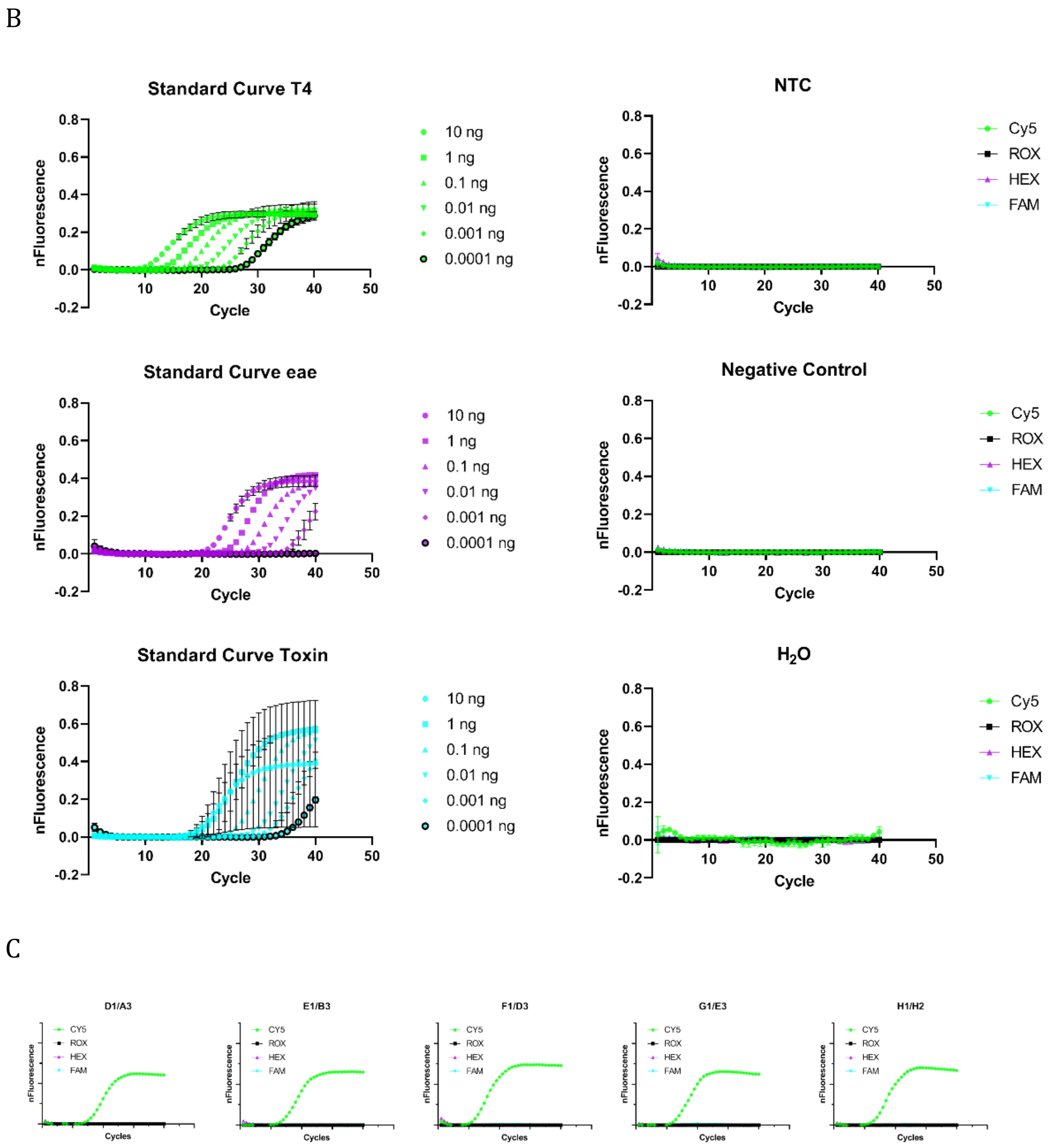
Primer_name	Sequence	Length	GC content	T_m	Product Length
E.coli_head_fow	TAACGGTAACTCCGCAGAATG	21	47.6	62	116
E.coli_head_rev	GATCGCGTAGTCGCTGTAAATA	22	45.5	62	
E.coli_eaeA_fow	GCTAAAGCGGATGGCATAGA	20	50	62	107
E.coli_eaeA_rev	GCAGTCCCGGATACAATACTAAA	23	43.5	62	
E.coli_toxin_fow_set2	GGCCCGCATCCAGTTATG	18	61.1	63	85
E.coli_toxin_rev_set2	GCGAGTGACGGCTTTGT	17	58.8	62	
E.coli_T4_head_fow	GGTACGTCGTGCTATTCCTAAC	22	50	62	117
E.coli_T4_head_rev	GCCACTGGGTCTTTACCATATAC	23	47.8	62	
E.coli_T4_eae_fow	CTCAGGCTAATGTCCCTGTAAC	22	50	62	89
E.coli_T4_eae_rev	CTTACCGTTACCATCCGTTCTG	22	50	62	
E.coli_T4_toxin_fow	TCAACACAGTATATCCGAAGGC	22	45.5	62	100
E.coli_T4_toxin_rev	GTGACGGCTTTGTAGTCCTT	20	50	62	
E.coli_T7_head_fow	CACGTCTTCCCTGCCAATAA	20	50	62	104
E.coli_T7_head_rev	CGCAGCTTAACAGTACCTACC	21	52.4	62	
E.coli_T7_eaeA_fow	GCTAAAGCGGATGGCATAGA	20	50	62	107
E.coli_T7_eaeA_rev	GCAGTCCCGGATACAATACTAAA	23	43.5	62	
E.coli_T7_toxin_fow	GGCCCGCATCCAGTTATG	18	61.1	63	85
E.coli_T7_toxin_rev	GCGAGTGACGGCTTTGT	17	58.8	62	
E.coli_T1_head_fow	TCGACGCGGTACAAACTAATATC	23	43.5	62	101
E.coli_T1_head_rev	GAACTCCTGTTCGGCATCAA	20	50	62	
E.coli_T4_head_fow	GGTACGTCGTGCTATTCCTAAC	22	50	62	117
E.coli_T4_head_rev	GCCACTGGGTCTTTACCATATAC	23	47.8	62	
E.coli_T7_head_fow	CACGTCTTCCCTGCCAATAA	20	50	62	104
E.coli_T7_head_rev	CGCAGCTTAACAGTACCTACC	21	52.4	62	

Supplementary Table 2_Probes for qPCR

Probe_Name	Sequence	Fluorophore / Quencher	Length	GC %	T_m
E.coli_head_probe	AAGGACGTTGTTGTCGGATCGTGT	Cy5 - BHQ3	24	50	68
E.coli_eaeA_probe	AATGGTGTAGCTCAGGCTAATGTCCC	HEX -BHQ1	26	50	68
E.coli_toxin_probe_set2	CATCGTGCATATGGTGCGCAACAG	FAM - BHQ1	24	54	68
E.coli_T4_head_probe	CGGCTGAACACCACAAATATCGAAAGC	Cy5 - BHQ3	27	48	68
E.coli_T4_eae_probe	CCCAAGAGTTGCAGTCCCGGATAC	HEX -BHQ1	24	58	68
E.coli_T4_toxin_probe	CATCGTGCATATGGTGCGCAACAG	FAM - BHQ1	24	54	68
E.coli_T7_head_probe	AGGTGAGGGTAATGTCAAGGTTGCT	Cy5 - BHQ3	25	48	68
E.coli_T7_eaeA_probe	AATGGTGTAGCTCAGGCTAATGTCCC	HEX -BHQ1	26	50	68
E.coli_T7_toxin_probe	CATCGTGCATATGGTGCGCAACAG	FAM - BHQ1	24	54	68
E.coli_T1_head_probe	AAAGGTCGTGCGGGAATTGCTAAA	Cy5 - BHQ3	24	46	67
E.coli_T4_head_probe	CGGCTGAACACCACAAATATCGAAAGC	HEX -BHQ1	27	48	68
E.coli_T7_head_probe	AGGTGAGGGTAATGTCAAGGTTGCT	FAM - BHQ1	25	48	68

A



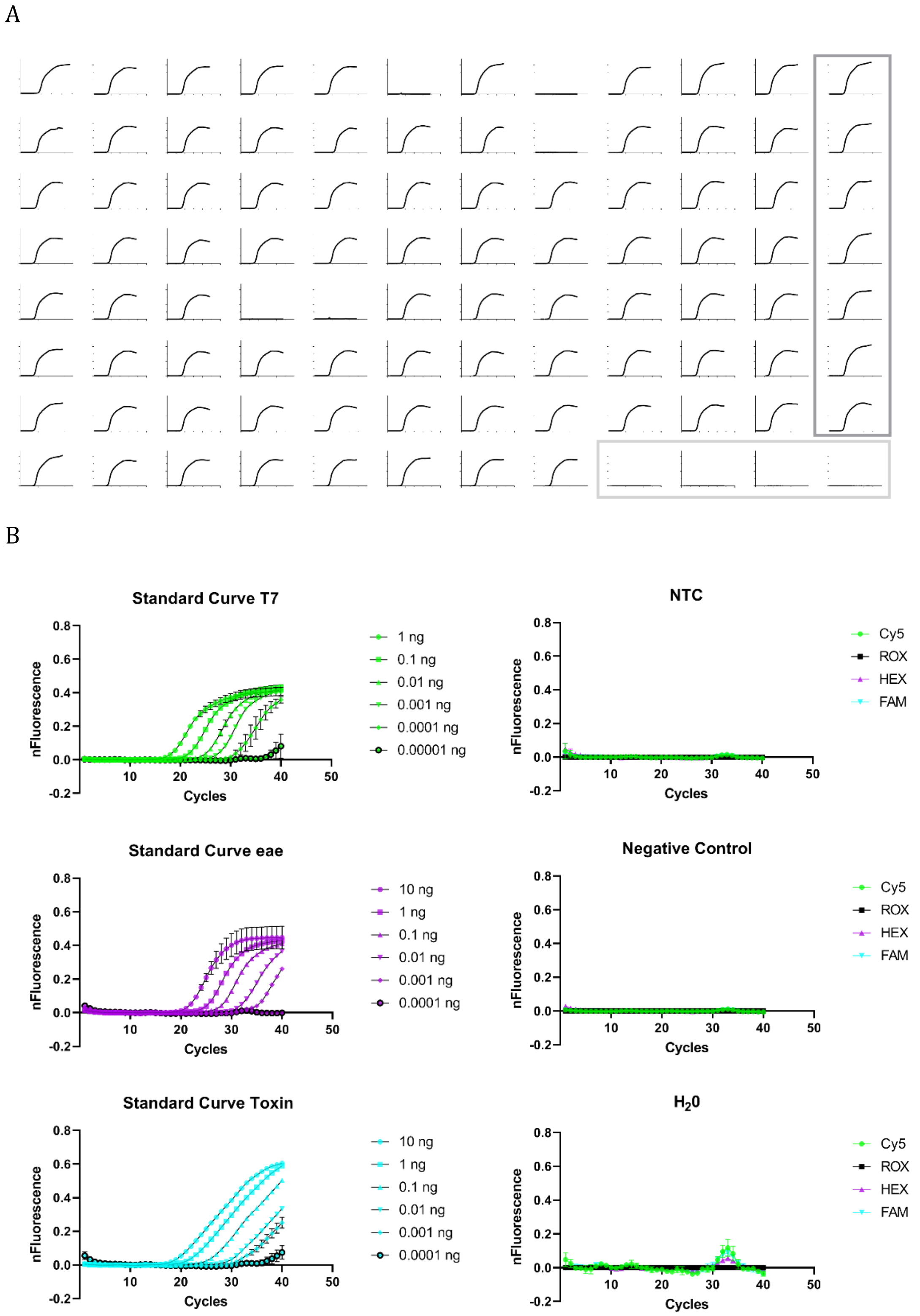


Supplementary Figure 1_Proof of Concept Results T4

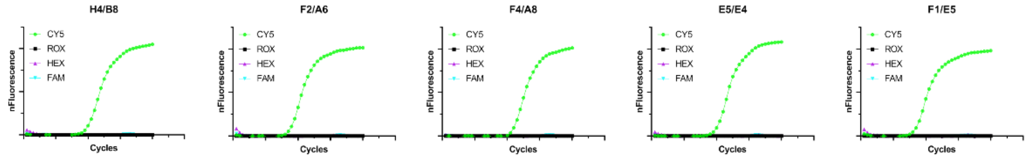
A_Infection Dynamics were monitored overnight. Bacterial-only controls are marked with a dark grey square, whereas media-only controls are with light grey.

B_On the left, qPCR controls are shown from T4, the toxin and the virulence factors with the amount of cycles on the x-axis and the normalized fluorescence on the y-axis. Six concentrations were tested from 10 ng to 1 pg. On the right, three negative controls are shown. On the top, there is the no-template control; in the middle, the negative control; and on the bottom, water only.

C_Five representatives of qPCR amplification of target samples. There are no amplifications from toxins or, virulence factors or cross-contaminations. All samples amplify around 15 cycles.


A
B
Standard Curve T7
nFluorescence
Cycles
1 ng
0.1 ng
0.01 ng
0.001 ng
0.0001 ng
0.00001 ng
NTC
Cy5
ROX
HEX
FAM
Standard Curve eae
10 ng
1 ng
0.1 ng
0.01 ng
0.001 ng
0.0001 ng
Negative Control
Standard Curve Toxin
H_20
0.8
0.6
0.4
0.2
0.0
-0.2
10
20
30
40
50

C



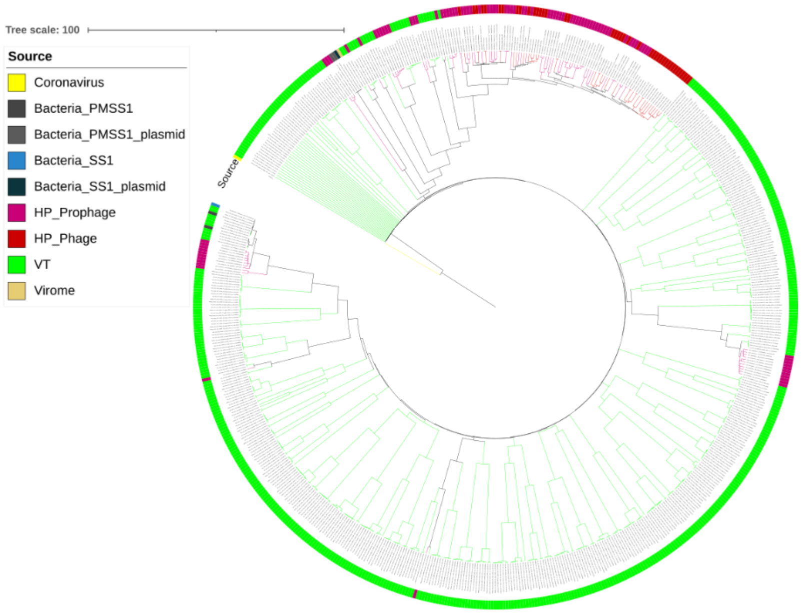
Supplementary Figure 2_Proof of Concept Results T7

A_Infection Dynamics were monitored overnight. Bacterial-only controls are marked with a dark grey square, whereas media-only controls are with light grey.

B_On the left, qPCR controls are shown from T7, the toxin and the virulence factors with the amount of cycles on the x-axis and the normalized fluorescence on the y-axis. Six concentrations were tested from 10 ng to 1 pg. On the right, three negative controls are shown. On the top, there is the no-template control; in the middle, the negative control; and on the bottom, water only.

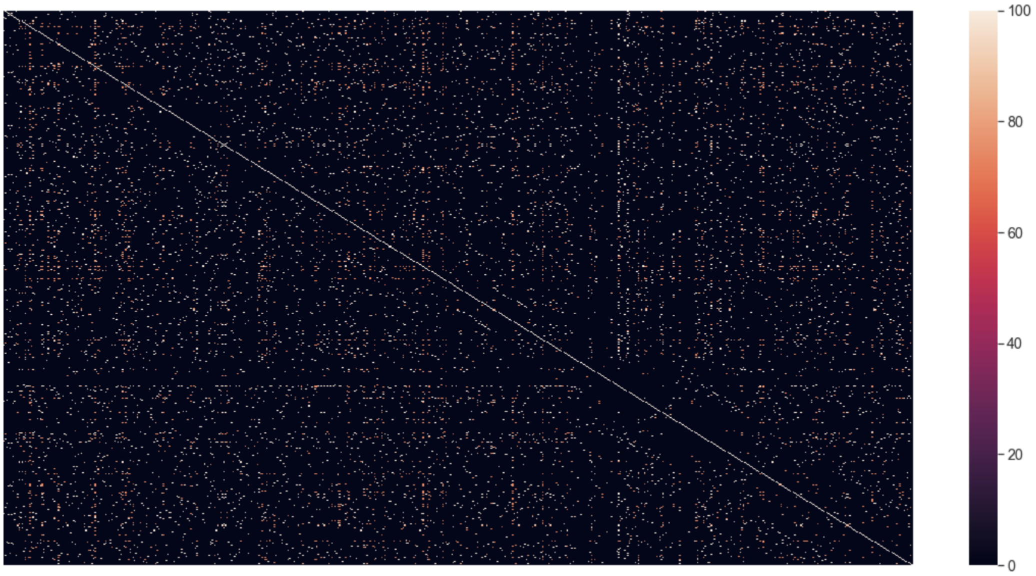
C_Five representatives of qPCR amplification of target samples. There are no amplifications from toxins or virulence factors or cross-contaminations. All samples amplify around 15 cycles.

6.1.2 *Helicobacter pylori* Phages and Their Diversity



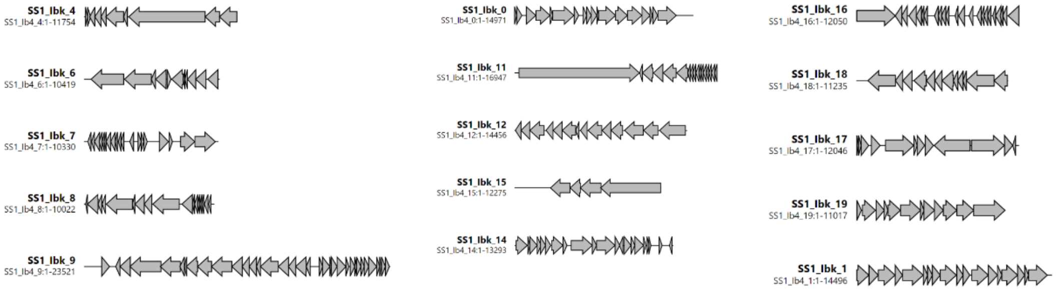
Supplementary Figure 3_Phylogenetic Tree From Viral-Tagged Contigs and Public Database Sequences

A phylogenetic tree with nine different sources presented as a re-rooted tree

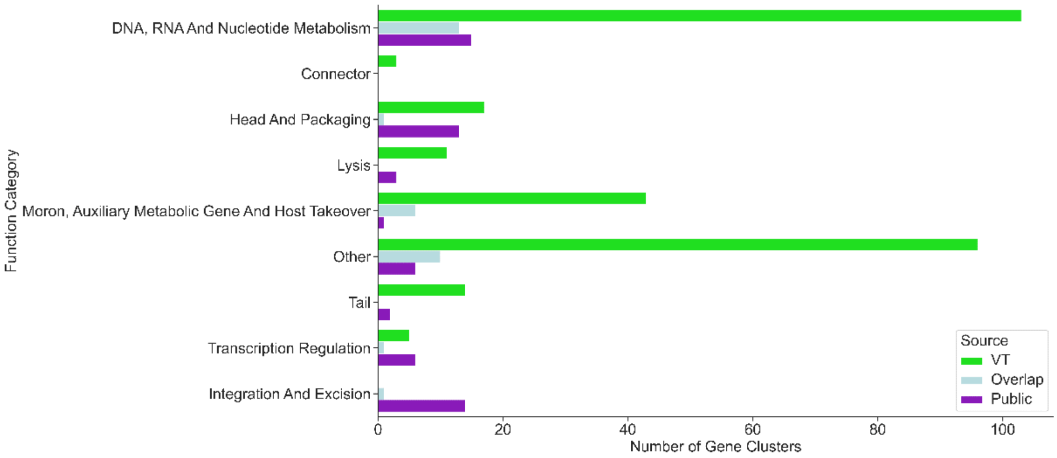


Supplementary Figure 4_ANI Comparison Heatmap

The heatmap shows the pairwise average nucleotide identity from each contig compared to each contig. Row and columns represent contigs, and the color code is their similarity. Black is zero similarity and white 100 % identity. The heatmap shows that assembled contigs are highly diverse and d



Supplementary Figure 5_ Highly Potential Phages From Category 2



Supplementary Figure 6_Gene Clusters

The graph presents the number of genes in absolute numbers for different gene functions from the in-house database (purple) and viral-tagged samples (VT, green). If genes overlapped in the groups, they were presented in blue.

6.2 List of Clinical Trials

Supplementary Table 3_List of Clinical Trials

#	NCT Number	Study Title	Study Status	Conditions	Interventions	Enrollment	Start Date	Completion Date
1	NCT05498363	Bacteriophage Therapy of Difficult-to-treat Infections	COMPLETED	Bacterial Infections	BIOLOGICAL: Bacteriophage therapy	100	01. Jan. 2008	31.DEC.2021
2	NCT04682964	Bacteriophage Therapy in Tonsillitis	ACTIVE_NOT_RECRUITING	Acute Tonsillitis	DRUG: Nebulizer inhalation irrigation of the mucous membranes of the tonsils with a bacteriophage.	128	02.OCT.2020	31.DEC.2024
3	NCT04287478	Bacteriophage Therapy in Patients With Urinary Tract Infections	TERMINATED	Urinary Tract Infection Bacterial	BIOLOGICAL: Bacteriophage Therapy	1	09.DEC.2020	28. Feb. 2023
4	NCT04596319	Ph 1/2 Study Evaluating Safety and Tolerability of Inhaled AP-PA02 in Subjects With Chronic Pseudomonas Aeruginosa Lung Infections and Cystic Fibrosis	COMPLETED	Cystic Fibrosis | Pseudomonas Aeruginosa | Pseudomonas | Lung Infection | Lung Infection Pseudomonal	BIOLOGICAL: AP-PA02 | OTHER: Placebo	29	22.DEC.2020	14.DEC.2022
5	NCT04803708	Bacteriophage Therapy TP-102 in Diabetic Foot Ulcers	COMPLETED	Diabetic Foot Ulcer | Pseudomonas Aeruginosa Infection | Staphylococcus Aureus Infection | Acinetobacter Infection	BIOLOGICAL: TP-102	20	22.MAR.2021	05. Sep. 2022
6	NCT04684641	CYstic Fibrosis bacterioPHage Study at Yale (CYPHY)	COMPLETED	Cystic Fibrosis	DRUG: Standard Dose YPT-01 | OTHER: Placebo	8	29.MAR.2021	22. Jun. 2023
7	NCT06368388	Bacteriophage Therapy for Difficult-to-treat Infections: the Implementation of a Multidisciplinary Phage Task Force	RECRUITING	Musculoskeletal Infection | Chronic Rhinosinusitis (Diagnosis) | Sepsis | Pulmonary Infection | Hidradenitis Suppurativa	OTHER: Prospective data collection | OTHER: Prospective data collection	50	01. Jun. 2021	01. Jun. 2025
8	NCT05177107	Bacteriophage Therapy in Patients With Diabetic Foot Osteomyelitis	RECRUITING	Osteomyelitis | Diabetic Foot Osteomyelitis	BIOLOGICAL: Phage Therapy | OTHER: Placebo	126	24. Nov. 2021	DEC.2024
9	NCT05184764	Study Evaluating Safety, Tolerability, and Efficacy of Intravenous AP-SA02 in Subjects With S. Aureus Bacteremia	RECRUITING	Bacteremia | Staphylococcus Aureus | Staphylococcus Aureus Bacteremia | Bacteremia Staph | Bacteremia Due to Staphylococcus Aureus	BIOLOGICAL: AP-SA02 | OTHER: Placebo	50	26. Apr. 2022	MAR.2025
10	NCT04650607	Phage Safety Cohort Study	RECRUITING	Prosthetic Joint Infection | Severe Infection	OTHER: Adverse event after injection of phages	100	09.MAY.2022	09.MAY.2028
11	NCT02664740	Standard Treatment Associated With Phage Therapy Versus Placebo for Diabetic Foot Ulcers Infected by S. Aureus	UNKNOWN	Diabetic Foot | Staphylococcal Infections	DRUG: Topical anti-Staphylococcus bacteriophage therapy | DRUG: Topical placebo corresponding to anti-Staphylococcus bacteriophage therapy	60	01. Jun. 2022	Aug. 2024
12	NCT05369104	Phage Therapy in Prosthetic Joint Infection Due to Staphylococcus Aureus Treated With DAIR.	RECRUITING	Infection of Total Hip Joint Prosthesis | Infection of Total Knee Joint Prosthesis	BIOLOGICAL: Anti-Staphylococcus aureus Bacteriophages	64	15. Jun. 2022	16. Jun 25
13	NCT05010577	Nebulized Bacteriophage Therapy in Cystic Fibrosis Patients With Chronic Pseudomonas Aeruginosa Pulmonary Infection	ACTIVE_NOT_RECRUITING	Chronic Pseudomonas Aeruginosa Infection | Cystic Fibrosis	DRUG: BX004-A | DRUG: Placebo	32	21. Jun. 2022	MAR. 2024
14	NCT05488340	A Study of LBP-EC01 in the Treatment of Acute Uncomplicated UTI Caused by Drug Resistant E. Coli (ELIMINATE Trial)	RECRUITING	Urinary Tract Infections	DRUG: LBP-EC01 0.1 x IV dose | DRUG: LBP-EC01 0.01x IV Dose | DRUG: LBP-EC01 IV Infusion Dose | DRUG: Placebo | DRUG: LBP-EC01 | DRUG: TMP/SMX	318	13. Jul. 2022	DEC. 2025
15	NCT05453578	A Phase 1b/2 Trial of the Safety and Microbiological Activity of Bacteriophage Therapy in Cystic Fibrosis Subjects Colonized With Pseudomonas Aeruginosa	RECRUITING	Bacterial Disease Carrier | Cystic Fibrosis	OTHER: Placebo | BIOLOGICAL: WRAIR-PAM-CF1	72	03.OCT.2022	31.DEC.2024
16	NCT05616221	Study to Evaluate the Safety, Phage Kinetics, and Efficacy of Inhaled AP-PA02 in Subjects With Non-Cystic Fibrosis Bronchiectasis and Chronic Pulmonary Pseudomonas Aeruginosa Infection	COMPLETED	Non-cystic Fibrosis Bronchiectasis | Pseudomonas Aeruginosa | Lung Infection	BIOLOGICAL: AP-PA02 | OTHER: Placebo	48	10. Jan. 2023	08. Aug. 2024
17	NCT05883995	Avis-PHAGEinLYON Clinic	RECRUITING	Bone Infection	OTHER: Phage therapy requests	1500	01. Feb. 2023	01. Feb. 2029
18	NCT06185920	PHAGEinLYON Clinic Cohort Study: a Descriptive Study of Severe Infections Treated With Phage Therapy at the HCL.	RECRUITING	Severe Infection	OTHER: Description of severe infection	250	01. Feb. 2023	01. Feb. 1934
19	NCT05269134	Bacteriophage Therapy in Patients With Prosthetic Joint Infections (PJI)	WITHDRAWN	Prosthetic Joint Infection	DRUG: Bacteriophage | DRUG: Placebo	0	27.MAR.2023	Jul. 2027
20	NCT05537519	Phage Therapy for the Treatment of Urinary Tract Infection	ACTIVE_NOT_RECRUITING	Recurrent Urinary Tract Infection	BIOLOGICAL: Phage Therapy	1	01.MAY.2023	30. Jun. 2024
21	NCT05973721	Clinical Study of Phage Therapy for Chronic Constipation Efficacy and Safety	RECRUITING	Pib Specific Phage | Intractable Constipation	BIOLOGICAL: phage	50	01. Jul. 2023	01.OCT.2023
22	NCT05967130	Treatment Chronic UTI Post Kidney Transplant	RECRUITING	Urinary Tract Infections | Transplant-Related Disorder	BIOLOGICAL: Phage	20	01. Jul. 2023	01. Jul. 2027
23	NCT06319235	Clinical Trial to Demonstrate the Safety and Efficacy of DUOFAGÂ®	RECRUITING	Surgical Site Infection | Staphylococcus Aureus Infection | Pseudomonas Aeruginosa Infection | Bacterial Infections | Surgical Wound Infection	DRUG: IMP | DRUG: Placebo	52	27.OCT.2023	31.DEC.2025
24	NCT05272579	PrePhage - Faecal Bacteriophage Transfer for Enhanced Gastrointestinal Tract Maturation in Preterm Infants	RECRUITING	Necrotizing Enterocolitis | Microbial Substitution	OTHER: Fecal Filtrate Transfer | OTHER: Placebo	20	07. Nov. 2023	31.DEC.2024
25	NCT05948592	Bacteriophage Therapy TP-102 in Patients With Diabetic Foot Infection	RECRUITING	Diabetic Foot Infection	BIOLOGICAL: TP-102 | OTHER: Placebo	80	08. Nov. 2023	31.DEC.2024
26	NCT06409819	Phage Therapy for Recurrent UTIs in Kidney Transplant Recipients	NOT_YET_RECRUITING	Urinary Tract Infection, Recurrent	DRUG: phage therapy | DRUG: control	32	01. Jun. 2024	30. Jun. 2027
27	NCT00945087	Experimental Phage Therapy of Bacterial Infections	UNKNOWN	Bacterial Infections	OTHER: Bacteriophage preparation		DEC. 2005	
28	NCT00663091	A Prospective, Randomized, Double-Blind Controlled Study of WPP-201 for the Safety and Efficacy of Treatment of Venous Leg Ulcers	COMPLETED	Venous Leg Ulcers	DRUG: Bacteriophage | DRUG: WPP-201 Bacteriophage | DRUG: Bacteriophages	64	Sep. 2006	MAY.2008
29	NCT00937274	Antibacterial Treatment Against Diarrhea in Oral Rehydration Solution	TERMINATED	Diarrhea	OTHER: T4 phage cocktail test | OTHER: Commercial T4 phage cocktail | OTHER: standard oral rehydration solution (ORS)	120	Aug. 2009	Jan. 2013
30	NCT02116010	Evaluation of Phage Therapy for the Treatment of Escherichia Coli and Pseudomonas Aeruginosa Wound Infections in Burned Patients	UNKNOWN	Wound Infection	DRUG: E. coli Phages cocktail | DRUG: Standard of care : Silver Sulfadiazine | DRUG: P. Aeruginosa, Phages cocktail		Jul. 2015	
31	NCT04323475	Phage Therapy for the Prevention and Treatment of Wound Infections in Burned Patients	UNKNOWN	Wound Infection	BIOLOGICAL: Bacteriophage cocktail spray | DRUG: Xeroform	12	Jan. 2022	Aug. 2023
32	NCT04815798	Phage Therapy for the Prevention and Treatment of Pressure Ulcers.	UNKNOWN	Pressure Ulcer	COMBINATION_PRODUCT: Bacteriophage-loaded Microcapsule Spray | COMBINATION_PRODUCT: Placebo | PROCEDURE: Standard of Care	69	Jan. 2022	DEC.2023
33	NCT05269121	Bacteriophage Therapy in First Time Chronic Prosthetic Joint Infections	WITHDRAWN	Prosthetic Joint Infection | Bacterial Infections	BIOLOGICAL: Phage Therapy	0	Sep. 2022	Nov. 2024
34	NCT04787250	Bacteriophage Therapy in Patients With Prosthetic Joint Infections	WITHDRAWN	Prosthetic Joint Infection	BIOLOGICAL: Phage Therapy | PROCEDURE: Two-Stage Exchange Arthroplasty	0	OCT.2022	MAY.2024
35	NCT06456424	Bacteriophage Therapy for Methicillin-Sensitive Staphylococcus Aureus Prosthetic Joint Infection	NOT_YET_RECRUITING	Prosthetic Joint Infections of Hip | Staphylococcus Aureus Infection	BIOLOGICAL: Phage therapy	1	Jun. 2024	Jun. 2025
36	NCT06370598	Phase 1/2a to Assess the Safety and Tolerability of TP-122A for the Treatment of Ventilator-Associated Pneumonia	NOT_YET_RECRUITING	Pneumonia, Ventilator-Associated	BIOLOGICAL: TP-122A	15	Sep. 2024	Jun. 2025
37	NCT04636554	Personalized Phage Treatment in Covid-19 Patients With Bacterial Co-Infections Microbials for Pneumonia or Bacteremia/Septicemia	NO_LONGER_AVAILABLE	Covid19 | Bacteremia | Septicemia | Acinetobacter Baumannii Infection | Pseudomonas Aeruginosa Infection | Staph Aureus Infection	OTHER: Phage Therapy			
38	NCT06262282	Mycobacteriophage Treatment of Non-tuberculosis Mycobacteria	ENROLLING_BY_INVITATION	Cystic Fibrosis | Nontuberculous Mycobacterial Lung Disease | Nontuberculous Mycobacterium Infection | Mycobacterium Infections | Mycobacterium				
39	NCT03395743	Individual Patient Expanded Access for AB-PA01, an Investigational Anti-Pseudomonas Aeruginosa Bacteriophage Therapeutic	NO_LONGER_AVAILABLE		BIOLOGICAL: AB-PA01			
40	NCT03395769	Individual Patient Expanded Access for AB-SA01, an Investigational Anti-Staphylococcus Aureus Bacteriophage Therapeutic	NO_LONGER_AVAILABLE		BIOLOGICAL: AB-SA01			
41	NCT05590195	Effect of PreforProÂ® on Urinary and Vaginal Health	NOT_YET_RECRUITING	Bacterial Vaginosis | Bacterial Infections | Bacterial Vaginosis &#X7C				

List of Abbreviations

Abbreviation	Explanation
µL	Microliter
µm	Micrometer
AMG	Auxiliary Metabolic Gene
ANI	Average Nucleotide Identity
AUC	Area Under the Curve
BD	Becton, Dickinson and Company
bp	Basepair
BVBRC	Bacterial and Viral Bioinformatics Resource Center
CagA	Cytotoxin-Associated Gene A
cagPAI	Cytotoxin-Associated Gene Pathogenicity Island
cfu/mL	Colony-Forming Unit per Milliliter
CRC	Colorectal Cancer
CRCA	Colorectal Cancer in Advanced Stadium
CRCE	Colorectal Cancer in Early Stadium
CRISPR cas9	Clustered Regularly Interspaced Palindromic Repeats
CT	Computed Tomography
DCC	Deleted in Colorectal Cancer Gene
(ds))DNA	(double-stranded) Deoxyribonucleic Acid
EMA	European Medicines Agency
ESKAPE	*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter spp*
FDA	Food and Drug Administration
FMT	Fecal Microbiota Transplantation
FSC	Forward Scatter
FVT	Fecal Viral Transplantation
GABA	Gamma-aminobutyric acid
GB	Gigabyte
GC-content	Guanine-Cytosine Content
GPD	Gut Phage Database
GI	Gastrointestinal Tract
GMP	Good Manufacturing Practice
***H. pylori* PMSS1**	*Helicobacter pylori* Pre-Mouse Sidney Strain
***H. pylori* SS1**	*Helicobacter pylori* Sidney Strain
HIV / HI virus	Human Immunodeficiency Virus
HP	*Helicobacter pylori*
IBD	Inflammatory Bowel Disease
ICTV	International Committee on Taxonomy of Viruses
iPHoP	Integrated Phage-Host Prediction
kpb	kilobasepairs
LB	Luria Bertani Broth

LCB	Local Colinear Blocks
LPS	Lipopolysaccharide
LT Toxin	Heat-labile Toxin
MALDI-TOFMS	Matrix-assisted laser desorption/ionization-time-of-flight mass spectrometry
MALT	Mucosa-associated lymphoid tissue lymphoma
Mbp	Megabasepair
MOI	Multiplicity of Infection
MWCO	Molecular weight Cutoff
NAD	Nicotinamide Adenine Dinucleotide
NCBI	National Center for Biotechnology Information
nm	Nanometer
nM	Nanomolar
NMR	Nuclear Magnetic Resonance
NT	Nucleotide Database
NTA	Nanoparticle Tracking Analysis
NTC	No Template Control
ORF	Open Reading Frame
PBS	Phosphate Buffered Saline
PCA	Principal Component Analysis
PET/CT	Positron Emission Tomography Computed Tomography
PFU/mL	Plaque Forming Unit per Milliliter
pg	Picogram
PHROG	Prokaryotic Virus Remote Homologous Groups
qPCR	Quantitative Polymerase Chain Reaction
rcf	Relative Centrifugal Force
(ss) RNA	(single-stranded) Ribonucleic Acid
rRNA	Ribosomal Ribonucleic Acid
SCFA	Short-Chain Fatty Acid
Spp	Species
SSC	Side Scatter
T4SS	Type 4 Secretion System
UC	Ulcerative Colitis
UK	United Kingdom
US / USA	United States of America
UV	Ultra-Violet
VacA	Vacuolating cytotoxin A
VC	Viral Clusters
VLP	Virus-Like-Particle
vOTU	Viral Operational Taxonomic Unit
VT	Viral Tagging
WHO	World Health Organization

List of Figures

List of Tables

Bibliography

[1] M. Lobanovska and G. Pilla, “Penicillin’s Discovery and Antibiotic Resistance: Lessons for the Future?,” *Yale J. Biol. Med.*, vol. 90, no. 1, pp. 135–145, Mar. 2017.

[2] G. Mancuso, A. Midiri, E. Gerace, and C. Biondo, “Bacterial Antibiotic Resistance: The Most Critical Pathogens,” *Pathogens*, vol. 10, no. 10, p. 1310, Oct. 2021, doi: 10.3390/pathogens10101310.

[3] A. Górski, R. Międzybrodzki, G. Węgrzyn, E. Jończyk-Matysiak, J. Borysowski, and B. Weber-Dąbrowska, “Phage therapy: Current status and perspectives,” *Med. Res. Rev.*, vol. 40, no. 1, pp. 459–463, 2020, doi: 10.1002/med.21593.

[4] E. M. Darby *et al.*, “Molecular mechanisms of antibiotic resistance revisited,” *Nat. Rev. Microbiol.*, vol. 21, no. 5, Art. no. 5, May 2023, doi: 10.1038/s41579-022-00820-y.

[5] C. J. Murray *et al.*, “Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis,” *Lancet Lond. Engl.*, vol. 399, no. 10325, pp. 629–655, Feb. 2022, doi: 10.1016/S0140-6736(21)02724-0.

[6] S. A. Strathdee, G. F. Hatfull, V. K. Mutalik, and R. T. Schooley, “Phage therapy: From biological mechanisms to future directions,” *Cell*, vol. 186, no. 1, pp. 17–31, Jan. 2023, doi: 10.1016/j.cell.2022.11.017.

[7] B. R. Lenneman, J. Fernbach, M. J. Loessner, T. K. Lu, and S. Kilcher, “Enhancing phage therapy through synthetic biology and genome engineering,” *Curr. Opin. Biotechnol.*, vol. 68, pp. 151–159, Apr. 2021, doi: 10.1016/J.COPBIO.2020.11.003.

[8] J. Romeyer Dherbey and F. Bertels, “The untapped potential of phage model systems as therapeutic agents,” *Virus Evol.*, vol. 10, no. 1, p. veae007, Jan. 2024, doi: 10.1093/ve/veae007.

[9] T. Thompson, “The staggering death toll of drug-resistant bacteria,” *Nature*, Jan. 2022, doi: 10.1038/d41586-022-00228-x.

[10] C. Willy *et al.*, “Phage Therapy in Germany—Update 2023,” *Viruses*, vol. 15, no. 2, Art. no. 2, Feb. 2023, doi: 10.3390/v15020588.

[11] M. Hutchings, A. Truman, and B. Wilkinson, “Antibiotics: past, present and future,” *Curr. Opin. Microbiol.*, vol. 51, pp. 72–80, Oct. 2019, doi: 10.1016/j.mib.2019.10.008.

[12] A. S. Nilsson, “Phage therapy—constraints and possibilities,” *Ups. J. Med. Sci.*, vol. 119, no. 2, pp. 192–198, May 2014, doi: 10.3109/03009734.2014.902878.

[13] G. P. C. Salmond and P. C. Fineran, “A century of the phage: past, present and future,” *Nat. Rev. Microbiol.*, vol. 13, no. 12, Art. no. 12, Dec. 2015, doi: 10.1038/nrmicro3564.

[14] M. R. Clokie, A. D. Millard, A. V. Letarov, and S. Heaphy, “Phages in nature,” *Bacteriophage*, vol. 1, no. 1, pp. 31–45, 2011, doi: 10.4161/bact.1.1.14942.

[15] G. Ofir and R. Sorek, “Contemporary Phage Biology: From Classic Models to New Insights,” *Cell*, vol. 172, no. 6, pp. 1260–1270, Mar. 2018, doi: 10.1016/j.cell.2017.10.045.

[16] F. L. Gordillo Altamirano and J. J. Barr, “Phage therapy in the postantibiotic era,” *Clin. Microbiol. Rev.*, vol. 32, no. 2, 2019, doi: 10.1128/CMR.00066-18.

[17] J. J. Dennehy and S. T. Abedon, “Phage Infection and Lysis,” in *Bacteriophages*, D. R. Harper, S. T. Abedon, B. H. Burrowes, and M. L. McConville, Eds., Cham: Springer International Publishing, 2021, pp. 341–383. doi: 10.1007/978-3-319-41986-2_53.

[18] A. Du Toit, “The language of phages,” *Nat. Rev. Microbiol.*, vol. 15, no. 3, Art. no. 3, Mar. 2017, doi: 10.1038/nrmicro.2017.8.

[19] R. Kongari *et al.*, “Phage spanins: diversity, topological dynamics and gene convergence,” *BMC Bioinformatics*, vol. 19, p. 326, Sep. 2018, doi: 10.1186/s12859-018-2342-8.

[20] M. Rajaure, J. Berry, R. Kongari, J. Cahill, and R. Young, “Membrane fusion during phage lysis,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 17, pp. 5497–5502, Apr. 2015, doi: 10.1073/pnas.1420588112.

[21] I.-N. Wang, D. L. Smith, and R. Young, “Holins: The Protein Clocks of Bacteriophage Infections,” *Annu. Rev. Microbiol.*, vol. 54, no. Volume 54, 2000, pp. 799–825, Oct. 2000, doi: 10.1146/annurev.micro.54.1.799.

[22] J. Łoś *et al.*, “Temperate Phages, Prophages, and Lysogeny,” in *Bacteriophages: Biology, Technology, Therapy*, D. R. Harper, S. T. Abedon, B. H. Burrowes, and M. L. McConville, Eds., Cham: Springer International Publishing, 2021, pp. 119–150. doi: 10.1007/978-3-319-41986-2_3.

[23] A. Chevallereau, B. J. Pons, S. van Houte, and E. R. Westra, “Interactions between bacterial and phage communities in natural environments,” *Nat. Rev. Microbiol.*, vol. 20, no. 1, pp. 49–62, Jan. 2022, doi: 10.1038/s41579-021-00602-y.

[24] M. B. Dion, F. Oechslin, and S. Moineau, “Phage diversity, genomics and phylogeny,” *Nat. Rev. Microbiol.*, vol. 18, no. 3, Art. no. 3, Mar. 2020, doi: 10.1038/s41579-019-0311-5.

[25] T. P. Honap, K. Sankaranarayanan, S. L. Schnorr, A. T. Ozga, C. Warinner, and C. M. L. Jr, “Biogeographic study of human gut-associated crAssphage suggests impacts from industrialization and recent expansion,” *PLOS ONE*, vol. 15, no. 1, p. e0226930, Jan. 2020, doi: 10.1371/journal.pone.0226930.

[26] M. K. Mirzaei and C. F. Maurice, “Ménage à trois in the human gut: Interactions between host, bacteria and phages,” *Nat. Rev. Microbiol.*, vol. 15, no. 7, pp. 397–408, Jun. 2017, doi: 10.1038/nrmicro.2017.30.

[27] A. Jurczak-Kurek *et al.*, “Biodiversity of bacteriophages: morphological and biological properties of a large group of phages isolated from urban sewage,” *Sci. Rep.*, vol. 6, no. 1, Art. no. 1, Oct. 2016, doi: 10.1038/srep34338.

[28] D. Turner, A. M. Kropinski, and E. M. Adriaenssens, “A Roadmap for Genome-Based Phage Taxonomy,” *Viruses*, vol. 13, no. 3, Art. no. 3, Mar. 2021, doi: 10.3390/v13030506.

[29] J. Bazan, I. Całkosiński, and A. Gamian, “Phage display—A powerful technique for immunotherapy,” *Hum. Vaccines Immunother.*, vol. 8, no. 12, pp. 1817–1828, Dec. 2012, doi: 10.4161/hv.21703.

[30] S. R. Casjens and R. W. Hendrix, “Bacteriophage lambda: early pioneer and still relevant,” *Virology*, vol. 0, pp. 310–330, May 2015, doi: 10.1016/j.virol.2015.02.010.

[31] K. N. Lam *et al.*, “Phage-delivered CRISPR-Cas9 for strain-specific depletion and genomic deletions in the gut microbiome,” *Cell Rep.*, vol. 37, no. 5, p. 109930, Nov. 2021, doi: 10.1016/j.celrep.2021.109930.

[32] J. C. Rees and K. J. Voorhees, “Simultaneous detection of two bacterial pathogens using bacteriophage amplification coupled with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry,” *Rapid Commun. Mass Spectrom.*, vol. 19, no. 19, pp. 2757–2761, 2005, doi: 10.1002/rcm.2107.

[33] P. Hyman, “Phages for phage therapy: Isolation, characterization, and host range breadth,” *Pharmaceuticals*, vol. 12, no. 1, Mar. 2019, doi: 10.3390/ph12010035.

[34] M. K. Mirzaei and A. S. Nilsson, “Isolation of phages for phage therapy: A comparison of spot tests and efficiency of plating analyses for determination of

host range and efficacy," *PLoS ONE*, vol. 10, no. 3, 2015, doi: 10.1371/journal.pone.0118557.

[35] "The promise of phages," *Nat. Biotechnol.*, vol. 41, no. 5, pp. 583–583, May 2023, doi: 10.1038/s41587-023-01807-7.

[36] B. Zalewska-Piątek, "Phage Therapy—Challenges, Opportunities and Future Prospects," *Pharmaceuticals*, vol. 16, no. 12, p. 1638, Nov. 2023, doi: 10.3390/ph16121638.

[37] R. Kebriaei *et al.*, "Optimization of Phage-Antibiotic Combinations against Staphylococcus aureus Biofilms," *Microbiol. Spectr.*, vol. 11, no. 3, pp. e04918-22, May 2023, doi: 10.1128/spectrum.04918-22.

[38] J. Fujiki, K. Nakamura, T. Nakamura, and H. Iwano, "Fitness Trade-Offs between Phage and Antibiotic Sensitivity in Phage-Resistant Variants: Molecular Action and Insights into Clinical Applications for Phage Therapy," *Int. J. Mol. Sci.*, vol. 24, no. 21, Art. no. 21, Jan. 2023, doi: 10.3390/ijms242115628.

[39] A. M. Segall, D. R. Roach, and S. A. Strathdee, "Stronger together? Perspectives on phage-antibiotic synergy in clinical applications of phage therapy," *Curr. Opin. Microbiol.*, vol. 51, pp. 46–50, Oct. 2019, doi: 10.1016/j.mib.2019.03.005.

[40] O. Nestegard *et al.*, "Helicobacter pylori resistance to antibiotics before and after treatment: Incidence of eradication failure," *PLOS ONE*, vol. 17, no. 4, p. e0265322, Apr. 2022, doi: 10.1371/journal.pone.0265322.

[41] E. Tshibangu-Kabamba and Y. Yamaoka, "Helicobacter pylori infection and antibiotic resistance — from biology to clinical implications," *Nat. Rev. Gastroenterol. Hepatol. 2021 189*, vol. 18, no. 9, pp. 613–629, May 2021, doi: 10.1038/s41575-021-00449-x.

[42] C. Sousa *et al.*, "Helicobacter pylori infection: from standard to alternative treatment strategies," *Crit. Rev. Microbiol.*, vol. 48, no. 3, pp. 376–396, May 2022, doi: 10.1080/1040841X.2021.1975643.

[43] S. Khosravi, R. Amini, M. R. Arabestani, S. S. Talebi, and F. A. Jalilian, "Isolation of a lytic bacteriophage for Helicobacter pylori," *Gene Rep.*, vol. 23, p. 101107, Jun. 2021, doi: 10.1016/j.genrep.2021.101107.

[44] P. Lehours *et al.*, "Genome Sequencing Reveals a Phage in Helicobacter pylori," *mBio*, vol. 2, no. 6, pp. e00239-11, Nov. 2011, doi: 10.1128/mBio.00239-11.

[45] A. B. Muñoz, J. Stepanian, A. A. Trespalacios, and F. F. Vale, "Bacteriophages of Helicobacter pylori," *Front. Microbiol.*, vol. 11, p. 12, Nov. 2020, doi: 10.3389/fmicb.2020.549084.

[46] J. L. Draper *et al.*, "Fallacy of the Unique Genome: Sequence Diversity within Single Helicobacter pylori Strains," *mBio*, vol. 8, no. 1, pp. e02321-16, Feb. 2017, doi: 10.1128/mBio.02321-16.

[47] V. Dyer *et al.*, "Genomic features of the Helicobacter pylori strain PMSS1 and its virulence attributes as deduced from its in vivo colonisation patterns," *Mol. Microbiol.*, vol. 110, no. 5, pp. 761–776, 2018, doi: 10.1111/mmi.14123.

[48] R. FitzGerald and S. M. Smith, "An Overview of Helicobacter pylori Infection," in *Helicobacter Pylori*, S. M. Smith, Ed., in Methods in Molecular Biology. , New York, NY: Springer US, 2021, pp. 1–14. doi: 10.1007/978-1-0716-1302-3_1.

[49] J.-M. Liou, P. Malfertheiner, S. I. Smith, E. M. El-Omar, and M.-S. Wu, "40 years after the discovery of Helicobacter pylori: towards elimination of H pylori for gastric cancer prevention," *The Lancet*, vol. 403, no. 10444, pp. 2570–2572, Jun. 2024, doi: 10.1016/S0140-6736(24)01171-1.

[50] F. F. Vale, A. P. A. Matos, P. Carvalho, and J. M. B. Vítor, “*Helicobacter pylori* Phage Screening,” *Microsc. Microanal.*, vol. 14, no. S3, pp. 150–151, Sep. 2008, doi: 10.1017/S1431927608089721.

[51] E. Heintschel. Von Heinegg, H. P. Nalik, and E. N. Y. 1993 Schmid, “Characterisation of a Helicobacter Pylori Phage (HP1),” *J. Med. Microbiol.*, vol. 38, no. 4, pp. 245–249, doi: 10.1099/00222615-38-4-245.

[52] E. N. Schmid, G. Von Recklinghausen, and R. Y. 1990 Ansorg, “Bacteriophages in Helicobacter (Campylobacter) Pylori,” *J. Med. Microbiol.*, vol. 32, no. 2, pp. 101–104, doi: 10.1099/00222615-32-2-101.

[53] J. Uchiyama *et al.*, “Complete Genome Sequences of Two Helicobacter pylori Bacteriophages Isolated from Japanese Patients,” *J. Virol.*, vol. 86, no. 20, pp. 11400–11401, Oct. 2012, doi: 10.1128/JVI.01767-12.

[54] C.-H. Luo, P.-Y. Chiou, C.-Y. Yang, and N.-T. Lin, “Genome, Integration, and Transduction of a Novel Temperate Phage of Helicobacter pylori,” *J. Virol.*, vol. 86, no. 16, pp. 8781–8792, Aug. 2012, doi: 10.1128/JVI.00446-12.

[55] E. Thursby and N. Juge, “Introduction to the human gut microbiota,” *Biochem. J.*, vol. 474, no. 11, pp. 1823–1836, Jun. 2017, doi: 10.1042/BCJ20160510.

[56] Y. Zhang, S. Sharma, L. Tom, Y.-T. Liao, and V. C. H. Wu, “Gut Phageome—An Insight into the Role and Impact of Gut Microbiome and Their Correlation with Mammal Health and Diseases,” *Microorganisms*, vol. 11, no. 10, p. 2454, Sep. 2023, doi: 10.3390/microorganisms11102454.

[57] M. Khan Mirzaei and C. F. Maurice, “The Mammalian Gut as a Matchmaker,” *Cell Host Microbe*, vol. 22, no. 6, pp. 726–727, Dec. 2017, doi: 10.1016/j.chom.2017.11.015.

[58] W. M. de Vos, H. Tilg, M. V. Hul, and P. D. Cani, “Gut microbiome and health: mechanistic insights,” *Gut*, vol. 71, no. 5, p. 1020, May 2022, doi: 10.1136/gutjnl-2021-326789.

[59] R. D. Hills, B. A. Pontefract, H. R. Mishcon, C. A. Black, S. C. Sutton, and C. R. Theberge, “Gut Microbiome: Profound Implications for Diet and Disease,” *Nutrients*, vol. 11, no. 7, p. 1613, Jul. 2019, doi: 10.3390/nu11071613.

[60] G. A. Kuziel and S. Rakoff-Nahoum, “The gut microbiome,” *Curr. Biol.*, vol. 32, no. 6, pp. R257–R264, Mar. 2022, doi: 10.1016/j.cub.2022.02.023.

[61] Y. Duan, R. Young, and B. Schnabl, “Bacteriophages and their potential for treatment of gastrointestinal diseases,” *Nat. Rev. Gastroenterol. Hepatol.*, vol. 19, no. 2, pp. 135–144, Feb. 2022, doi: 10.1038/s41575-021-00536-z.

[62] B. Duan *et al.*, “Colorectal Cancer: An Overview,” in *Gastrointestinal Cancers*, J. A. Morgado-Diaz, Ed., Brisbane (AU): Exon Publications, 2022. Accessed: Aug. 18, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK586003/

[63] M. Rebersek, “Gut microbiome and its role in colorectal cancer,” *BMC Cancer*, vol. 21, no. 1, p. 1325, Dec. 2021, doi: 10.1186/s12885-021-09054-2.

[64] Y. Liu, H. C.-H. Lau, W. Y. Cheng, and J. Yu, “Gut Microbiome in Colorectal Cancer: Clinical Diagnosis and Treatment,” *Genomics Proteomics Bioinformatics*, vol. 21, no. 1, pp. 84–96, Feb. 2023, doi: 10.1016/j.gpb.2022.07.002.

[65] J. R. Marchesi *et al.*, “Towards the Human Colorectal Cancer Microbiome,” *PLOS ONE*, vol. 6, no. 5, p. e20447, May 2011, doi: 10.1371/journal.pone.0020447.

[66] J.-W. Wang *et al.*, “Fecal microbiota transplantation: Review and update,” *J. Formos. Med. Assoc.*, vol. 118, pp. S23–S31, Mar. 2019, doi: 10.1016/j.jfma.2018.08.011.

[67] S. Porcari *et al.*, "Key determinants of success in fecal microbiota transplantation: From microbiome to clinic," *Cell Host Microbe*, vol. 31, no. 5, pp. 712–733, May 2023, doi: 10.1016/j.chom.2023.03.020.

[68] R. Yang, Z. Chen, and J. Cai, "Fecal microbiota transplantation: Emerging applications in autoimmune diseases," *J. Autoimmun.*, vol. 141, p. 103038, Dec. 2023, doi: 10.1016/j.jaut.2023.103038.

[69] C. Zeng *et al.*, "Fecal virome transplantation: A promising strategy for the treatment of metabolic diseases," *Biomed. Pharmacother.*, vol. 177, p. 117065, Aug. 2024, doi: 10.1016/j.biopha.2024.117065.

[70] Y. Yu, W. Wang, and F. Zhang, "The Next Generation Fecal Microbiota Transplantation: To Transplant Bacteria or Virome," *Adv. Sci.*, vol. 10, no. 35, p. 2301097, Nov. 2023, doi: 10.1002/advs.202301097.

[71] S. Shen, D. Huo, C. Ma, S. Jiang, and J. Zhang, "Expanding the Colorectal Cancer Biomarkers Based on the Human Gut Phageome," *Microbiol. Spectr.*, vol. 9, no. 3, pp. e00090-21, doi: 10.1128/Spectrum.00090-21.

[72] W. Zuo, S. Michail, and F. Sun, "Metagenomic Analyses of Multiple Gut Datasets Revealed the Association of Phage Signatures in Colorectal Cancer," *Front. Cell. Infect. Microbiol.*, vol. 12, Jun. 2022, doi: 10.3389/fcimb.2022.918010.

[73] S. Luo *et al.*, "Gut virome profiling identifies an association between temperate phages and colorectal cancer promoted by Helicobacter pylori infection," *Gut Microbes*, vol. 15, no. 2, p. 2257291, Dec. 2023, doi: 10.1080/19490976.2023.2257291.

[74] L. Deng, A. Gregory, S. Yilmaz, B. T. Poulos, P. Hugenholtz, and M. B. Sullivan, "Contrasting life strategies of viruses that infect photo- and heterotrophic bacteria, as revealed by viral tagging," *mBio*, vol. 3, no. 6, Dec. 2012, doi: 10.1128/mBio.00373-12.

[75] M. Unterer, M. K. Mirzaei, and L. Deng, "Targeted Single-Phage Isolation Reveals Phage-Dependent Heterogeneous Infection Dynamics." Accessed: Oct. 04, 2024. [Online]. Available: https://journals.asm.org/doi/epub/10.1128/spectrum.05149-22

[76] K. Kondo, M. Kawano, and M. Sugai, "Distribution of Antimicrobial Resistance and Virulence Genes within the Prophage-Associated Regions in Nosocomial Pathogens," *mSphere*, vol. 6, no. 4, pp. e00452-21, Jul. 2021, doi: 10.1128/mSphere.00452-21.

[77] K. V. Gernaey, M. C. M. Van Loosdrecht, M. Henze, M. Lind, and S. B. Jørgensen, "Activated sludge wastewater treatment plant modelling and simulation: state of the art," *Environ. Model. Softw.*, vol. 19, no. 9, pp. 763–783, Sep. 2004, doi: 10.1016/j.envsoft.2003.03.005.

[78] S. Chen, Y. Zhou, Y. Chen, and J. Gu, "fastp: an ultra-fast all-in-one FASTQ preprocessor," *Bioinformatics*, vol. 34, no. 17, pp. i884–i890, Sep. 2018, doi: 10.1093/bioinformatics/bty560.

[79] S. Nurk, D. Meleshko, A. Korobeynikov, and P. A. Pevzner, "metaSPAdes: a new versatile metagenomic assembler," *Genome Res.*, vol. 27, no. 5, pp. 824–834, May 2017, doi: 10.1101/gr.213959.116.

[80] B. J. Walker *et al.*, "Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement," *PLoS ONE*, vol. 9, no. 11, p. e112963, Nov. 2014, doi: 10.1371/journal.pone.0112963.

[81] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, Jul. 2014, doi: 10.1093/bioinformatics/btu153.

[82] J. R. Grant *et al.*, “Proksee: in-depth characterization and visualization of bacterial genomes,” *Nucleic Acids Res.*, vol. 51, no. W1, pp. W484–W492, Jul. 2023, doi: 10.1093/nar/gkad326.

[83] N. A. O’Leary *et al.*, “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D733–D745, Jan. 2016, doi: 10.1093/nar/gkv1189.

[84] S. Nayfach, A. P. Camargo, F. Schulz, E. Eloe-Fadrosh, S. Roux, and N. C. Kyrpides, “CheckV assesses the quality and completeness of metagenome-assembled viral genomes,” *Nat. Biotechnol.*, vol. 39, no. 5, pp. 578–585, May 2021, doi: 10.1038/s41587-020-00774-7.

[85] S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan, “VirSorter: mining viral signal from microbial genomic data,” *PeerJ*, vol. 3, p. e985, May 2015, doi: 10.7717/peerj.985.

[86] S. Roux *et al.*, “iPHoP: an integrated machine-learning framework to maximize host prediction for metagenome-assembled virus genomes,” *bioRxiv*, p. 2022.07.28.501908, Jan. 2022, doi: 10.1101/2022.07.28.501908.

[87] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006, doi: 10.1093/bioinformatics/btl158.

[88] C. L. M. Gilchrist and Y.-H. Chooi, “clinker & clustermap.js: automatic generation of gene cluster comparison figures,” *Bioinformatics*, vol. 37, no. 16, pp. 2473–2475, Aug. 2021, doi: 10.1093/bioinformatics/btab007.

[89] M. Shaffer *et al.*, “DRAM for distilling microbial metabolism to automate the curation of microbiome function,” *Nucleic Acids Res.*, vol. 48, no. 16, pp. 8883–8900, Sep. 2020, doi: 10.1093/nar/gkaa621.

[90] D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser, “Prodigal: prokaryotic gene recognition and translation initiation site identification,” *BMC Bioinformatics*, vol. 11, p. 119, Mar. 2010, doi: 10.1186/1471-2105-11-119.

[91] J. Mistry *et al.*, “Pfam: The protein families database in 2021,” *Nucleic Acids Res.*, vol. 49, no. D1, Aug. 2021, doi: 10.1093/nar/gkaa913.

[92] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[93] P. Terzian *et al.*, “PHROG: families of prokaryotic virus proteins clustered using remote homology,” *NAR Genomics Bioinforma.*, vol. 3, no. 3, p. lqab067, Sep. 2021, doi: 10.1093/nargab/lqab067.

[94] J. Mistry, R. D. Finn, S. R. Eddy, A. Bateman, and M. Punta, “Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions,” *Nucleic Acids Res.*, vol. 41, no. 12, p. e121, Jul. 2013, doi: 10.1093/nar/gkt263.

[95] M. Steinegger and J. Söding, “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets,” *Nat. Biotechnol.*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, doi: 10.1038/nbt.3988.

[96] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform,” *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/nar/gkf436.

[97] S. Capella-Gutiérrez, J. M. Silla-Martínez, and T. Gabaldón, “trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses,” *Bioinforma. Oxf. Engl.*, vol. 25, no. 15, pp. 1972–1973, Aug. 2009, doi: 10.1093/bioinformatics/btp348.

[98] B. Q. Minh *et al.*, “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era,” *Mol. Biol. Evol.*, vol. 37, no. 5, pp. 1530–1534, May 2020, doi: 10.1093/molbev/msaa015.

[99] I. Letunic and P. Bork, “Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation,” *Nucleic Acids Res.*, vol. 49, no. W1, pp. W293–W296, Jul. 2021, doi: 10.1093/nar/gkab301.

[100] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nat. Methods*, vol. 9, no. 4, pp. 357–359, Apr. 2012, doi: 10.1038/nmeth.1923.

[101] P. Danecek *et al.*, “Twelve years of SAMtools and BCFtools,” *GigaScience*, vol. 10, no. 2, p. giab008, Feb. 2021, doi: 10.1093/gigascience/giab008.

[102] A. Almeida *et al.*, “A unified catalog of 204,938 reference genomes from the human gut microbiome,” *Nat. Biotechnol.*, vol. 39, no. 1, pp. 105–114, Jan. 2021, doi: 10.1038/s41587-020-0603-3.

[103] A. P. Camargo *et al.*, “Identification of mobile genetic elements with geNomad,” *Nat. Biotechnol.*, vol. 42, no. 8, pp. 1303–1312, Aug. 2024, doi: 10.1038/s41587-023-01953-y.

[104] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes,” *Genome Res.*, vol. 25, no. 7, pp. 1043–1055, Jul. 2015, doi: 10.1101/gr.186072.114.

[105] A. Mitrofanov, O. S. Alkhnbashi, S. A. Shmakov, K. S. Makarova, E. V. Koonin, and R. Backofen, “CRISPRidentify: identification of CRISPR arrays using machine learning approach,” *Nucleic Acids Res.*, vol. 49, no. 4, pp. e20–e20, Feb. 2021, doi: 10.1093/nar/gkaa1158.

[106] M. Steinegger and J. Söding, “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets,” *Nat. Biotechnol.*, vol. 35, no. 11, pp. 1026–1028, Nov. 2017, doi: 10.1038/nbt.3988.

[107] D. E. Wood and S. L. Salzberg, “Kraken: ultrafast metagenomic sequence classification using exact alignments,” *Genome Biol.*, vol. 15, no. 3, p. R46, Mar. 2014, doi: 10.1186/gb-2014-15-3-r46.

[108] B. D. Ondov *et al.*, “Mash Screen: high-throughput sequence containment estimation for genome discovery,” *Genome Biol.*, vol. 20, no. 1, p. 232, Dec. 2019, doi: 10.1186/s13059-019-1841-x.

[109] L. F. Camarillo-Guerrero, A. Almeida, G. Rangel-Pineros, R. D. Finn, and T. D. Lawley, “Massive expansion of human gut bacteriophage diversity,” *Cell*, vol. 184, no. 4, pp. 1098-1109.e9, Feb. 2021, doi: 10.1016/j.cell.2021.01.029.

[110] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, “NCBI BLAST: a better web interface,” *Nucleic Acids Res.*, vol. 36, no. suppl_2, pp. W5–W9, Jul. 2008, doi: 10.1093/nar/gkn201.

[111] M. B. Dion, P.-L. Plante, E. Zufferey, S. A. Shah, J. Corbeil, and S. Moineau, “Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter,” *Nucleic Acids Res.*, vol. 49, no. 6, pp. 3127–3138, Apr. 2021, doi: 10.1093/nar/gkab133.

[112] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, “UpSet: Visualization of Intersecting Sets,” *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1983–1992, Dec. 2014, doi: 10.1109/TVCG.2014.2346248.

[113] S. Peschel, C. L. Müller, E. Von Mutius, A.-L. Boulesteix, and M. Depner, “NetCoMi: network construction and comparison for microbiome data in R,” *Brief. Bioinform.*, vol. 22, no. 4, p. bbaa290, Jul. 2021, doi: 10.1093/bib/bbaa290.

[114] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[115] M. Džunková, S. J. Low, J. N. Daly, L. Deng, C. Rinke, and P. Hugenholtz, “Defining the human gut host–phage network through single-cell viral tagging,” *Nat. Microbiol.*, vol. 4, no. 12, pp. 2192–2203, 2019, doi: 10.1038/s41564-019-0526-2.

[116] K. Kosznik-Kwaśnicka *et al.*, “Propagation, Purification, and Characterization of Bacteriophages for Phage Therapy,” in *Bacteriophages: Methods and Protocols*, E. Tumban, Ed., New York, NY: Springer US, 2024, pp. 357–400. doi: 10.1007/978-1-0716-3549-0_22.

[117] N. S. Olsen, N. B. Hendriksen, L. H. Hansen, and W. Kot, “A New High-throughput Screening (HiTS) Method for Phages - Enabling Crude Isolation and Fast Identification of Diverse Phages with Therapeutic Potential,” *bioRxiv*, p. 2020.03.27.011080, 2020, doi: 10.1101/2020.03.27.011080.

[118] A. Afrizal *et al.*, “Anaerobic single-cell dispensing facilitates the cultivation of human gut bacteria,” *Environ. Microbiol.*, vol. 24, no. 9, pp. 3861–3881, 2022, doi: 10.1111/1462-2920.15935.

[119] X. Yang *et al.*, “Genetic diversity of the intimin gene (eae) in non-O157 Shiga toxin-producing Escherichia coli strains in China,” *Sci. Rep.*, vol. 10, no. 1, p. 3275, Feb. 2020, doi: 10.1038/s41598-020-60225-w.

[120] J. Kaletta, C. Pickl, C. Griebler, A. Klingl, R. Kurmayer, and L. Deng, “A rigorous assessment and comparison of enumeration methods for environmental viruses,” *Sci. Rep.*, vol. 10, no. 1, p. 18625, Oct. 2020, doi: 10.1038/s41598-020-75490-y.

[121] H. Duyvejonck *et al.*, “Development of a qPCR platform for quantification of the five bacteriophages within bacteriophage cocktail 2 (BFC2),” *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019, doi: 10.1038/s41598-019-50461-0.

[122] D. Refardt, “Real-time quantitative PCR to discriminate and quantify lambdoid bacteriophages of Escherichia coli K-12,” *Bacteriophage*, vol. 2, no. 2, pp. 98–104, Apr. 2012, doi: 10.4161/bact.20092.

[123] I. Auzat, M. Ouldali, E. Jacquet, B. Fauler, T. Mielke, and P. Tavares, “Dual function of a highly conserved bacteriophage tail completion protein essential for bacteriophage infectivity,” *Commun. Biol.*, vol. 7, no. 1, pp. 1–12, May 2024, doi: 10.1038/s42003-024-06221-6.

[124] H. Georjon and A. Bernheim, “The highly diverse antiphage defence systems of bacteria,” *Nat. Rev. Microbiol.*, vol. 21, no. 10, pp. 686–700, Oct. 2023, doi: 10.1038/s41579-023-00934-x.

[125] K. Murtazalieva, A. Mu, A. Petrovskaya, and R. D. Finn, “The growing repertoire of phage anti-defence systems,” *Trends Microbiol.*, vol. 0, no. 0, Jun. 2024, doi: 10.1016/j.tim.2024.05.005.

[126] S. L. Xiaoqing WANG, “Defense and anti-defense mechanisms of bacteria and bacteriophages,” *J. Zhejiang Univ. Sci. B*, vol. 25, no. 3, pp. 181–196, Mar. 2024, doi: 10.1631/jzus.B2300101.

[127] P. Hyman and S. T. Abedon, “Bacteriophage: Overview☆,” in *Encyclopedia of Microbiology (Fourth Edition)*, T. M. Schmidt, Ed., Oxford: Academic Press, 2019, pp. 441–457. doi: 10.1016/B978-0-12-801238-3.02506-X.

[128] D. Bryan, A. El-Shibiny, Z. Hobbs, J. Porter, and E. M. Kutter, “Bacteriophage T4 infection of stationary phase E. coli: Life after log from a phage perspective,”

Front. Microbiol., vol. 7, no. SEP, p. 1391, Sep. 2016, doi: 10.3389/FMICB.2016.01391/BIBTEX.

[129] S. V. Avery, "Microbial cell individuality and the underlying sources of heterogeneity," *Nat. Rev. Microbiol.*, vol. 4, no. 8, pp. 577–587, Aug. 2006, doi: 10.1038/nrmicro1460.

[130] M. D. Roberts, N. L. Martin, and A. M. Kropinski, "The genome and proteome of coliphage T1," *Virology*, vol. 318, no. 1, pp. 245–266, Jan. 2004, doi: 10.1016/J.VIROL.2003.09.020.

[131] K. Šivec and A. Podgornik, "Determination of bacteriophage growth parameters under cultivating conditions," *Appl. Microbiol. Biotechnol.*, vol. 104, no. 20, pp. 8949–8960, Oct. 2020, doi: 10.1007/s00253-020-10866-8.

[132] C. McGoverin, J. Robertson, Y. Jonmohamadi, S. Swift, and F. Vanholsbeeck, "Species Dependence of SYTO 9 Staining of Bacteria," *Front. Microbiol.*, vol. 11, p. 545419, Sep. 2020, doi: 10.3389/fmicb.2020.545419.

[133] H. Drexler, "Bacteriophage T1," in *The Bacteriophages*, R. Calendar, Ed., Boston, MA: Springer US, 1988, pp. 235–258. doi: 10.1007/978-1-4684-5424-6_7.

[134] A. Rabinovitch, H. Hadas, M. Einav, Z. Melamed, and A. Zaritsky, "Model for Bacteriophage T4 Development in Escherichia coli," *J. Bacteriol.*, vol. 181, no. 5, pp. 1677–1683, Mar. 1999.

[135] S. T. Abedon, "Lysis of lysis-inhibited bacteriophage T4-infected cells.," *J. Bacteriol.*, vol. 174, no. 24, pp. 8073–8080, Dec. 1992.

[136] S. Sillankorva, P. Neubauer, and J. Azeredo, "Isolation and characterization of a T7-like lytic phage for Pseudomonas fluorescens," *BMC Biotechnol.*, vol. 8, no. 1, p. 80, Oct. 2008, doi: 10.1186/1472-6750-8-80.

[137] H. Yue, Y. Li, M. Yang, and C. Mao, "T7 Phage as an Emerging Nanobiomaterial with Genetically Tunable Target Specificity," *Adv. Sci.*, vol. 9, no. 4, p. 2103645, Dec. 2021, doi: 10.1002/advs.202103645.

[138] H. B. Jang *et al.*, "Viral tag and grow: a scalable approach to capture and characterize infectious virus–host pairs," *ISME Commun. 2022 21*, vol. 2, no. 1, pp. 1–11, Feb. 2022, doi: 10.1038/s43705-022-00093-9.

[139] R. I. Dascălu *et al.*, "Multidrug resistance in Helicobacter pylori infection," *Front. Microbiol.*, vol. 14, Feb. 2023, doi: 10.3389/fmicb.2023.1128497.

[140] V. E. Reyes, "Helicobacter pylori and Its Role in Gastric Cancer," *Microorganisms*, vol. 11, no. 5, p. 1312, May 2023, doi: 10.3390/microorganisms11051312.

[141] L. M. Brown, "Helicobacter pylori: epidemiology and routes of transmission," *Epidemiol. Rev.*, vol. 22, no. 2, pp. 283–297, 2000, doi: 10.1093/oxfordjournals.epirev.a018040.

[142] K. Stefano *et al.*, "Helicobacter pylori, transmission routes and recurrence of infection: state of the art," *Acta Bio Medica Atenei Parm.*, vol. 89, no. Suppl 8, pp. 72–76, 2018, doi: 10.23750/abm.v89i8-S.7947.

[143] X.-Y. Zhang, P.-Y. Zhang, and M. A. M. Aboul-Soud, "From inflammation to gastric cancer: Role of Helicobacter pylori," *Oncol. Lett.*, vol. 13, no. 2, pp. 543–548, Feb. 2017, doi: 10.3892/ol.2016.5506.

[144] "KA-03391 · NLM Customer Support Center." Accessed: Aug. 10, 2024. [Online]. Available: https://support.nlm.nih.gov/knowledgebase/article/KA-03391/en-us

[145] D. Falush, "The Remarkable Genetics of Helicobacter pylori," *mBio*, vol. 13, no. 6, pp. e02158-22, doi: 10.1128/mbio.02158-22.

[146] "Helicobacter pylori SS1 chromosome, complete genome." Mar. 05, 2024. Accessed: Aug. 10, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/NZ_CP009259.1

[147] "Helicobacter pylori strain PMSS1 complete genome." Feb. 10, 2017. Accessed: Aug. 10, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/nuccore/CP018823.1

[148] R. D. Olson *et al.*, "Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D678–D689, Jan. 2023, doi: 10.1093/nar/gkac1003.

[149] A. B. Muñoz, A. A. Trespalacios-Rangel, and F. F. Vale, "An American lineage of Helicobacter pylori prophages found in Colombia," *Helicobacter*, vol. 26, no. 2, p. e12779, 2021, doi: 10.1111/hel.12779.

[150] F. F. Vale *et al.*, "Genomic structure and insertion sites of Helicobacter pylori prophages from various geographical origins," *Sci. Rep.*, vol. 7, no. 1, p. 42471, Feb. 2017, doi: 10.1038/srep42471.

[151] C. H. Gauthier, S. G. Cresawn, and G. F. Hatfull, "PhaMMseqs: a new pipeline for constructing phage gene phamilies using MMseqs2," *G3 GenesGenomesGenetics*, vol. 12, no. 11, p. jkac233, Sep. 2022, doi: 10.1093/g3journal/jkac233.

[152] P. K. Montso, A. M. Kropinski, F. Mokoena, R. E. Pierneef, V. Mlambo, and C. N. Ateba, "Comparative genomics and proteomics analysis of phages infecting multi-drug resistant Escherichia coli O177 isolated from cattle faeces," *Sci. Rep.*, vol. 13, no. 1, p. 21426, Dec. 2023, doi: 10.1038/s41598-023-48788-w.

[153] H. S. Gweon *et al.*, "The impact of sequencing depth on the inferred taxonomic composition and AMR gene content of metagenomic samples," *Environ. Microbiome*, vol. 14, no. 1, p. 7, Oct. 2019, doi: 10.1186/s40793-019-0347-1.

[154] X.-Q. Luo *et al.*, "Viral community-wide auxiliary metabolic genes differ by lifestyles, habitats, and hosts," *Microbiome*, vol. 10, no. 1, p. 190, Nov. 2022, doi: 10.1186/s40168-022-01384-y.

[155] H. Yu *et al.*, "Genetic diversity of virus auxiliary metabolism genes associated with phosphorus metabolism in Napahai plateau wetland," *Sci. Rep.*, vol. 13, no. 1, p. 3250, Feb. 2023, doi: 10.1038/s41598-023-28488-1.

[156] C. Selvaraj and S. K. Singh, "Phage Protein Interactions in the Inhibition Mechanism of Bacterial Cell," in *Biocommunication of Phages*, G. Witzany, Ed., Cham: Springer International Publishing, 2020, pp. 121–142. doi: 10.1007/978-3-030-45885-0_6.

[157] A. S. A. Dowah and M. R. J. Clokie, "Review of the nature, diversity and structure of bacteriophage receptor binding proteins that target Gram-positive bacteria," *Biophys. Rev.*, vol. 10, no. 2, pp. 535–542, Jan. 2018, doi: 10.1007/s12551-017-0382-3.

[158] D. Rothschild-Rodriguez, M. Hedges, M. Kaplan, S. Karav, and F. L. Nobrega, "Phage-encoded carbohydrate-interacting proteins in the human gut," *Front. Microbiol.*, vol. 13, Jan. 2023, doi: 10.3389/fmicb.2022.1083208.

[159] S. V. Owen, R. Canals, N. Wenner, D. L. Hammarlöf, C. Kröger, and J. C. D. Hinton, "A window into lysogeny: revealing temperate phage biology with transcriptomics," *Microb. Genomics*, vol. 6, no. 2, p. e000330, Feb. 2020, doi: 10.1099/mgen.0.000330.

[160] D. Y. Lee, C. Bartels, K. McNair, R. A. Edwards, M. A. Swairjo, and A. Luque, "Predicting the capsid architecture of phages from metagenomic data," *Comput. Struct. Biotechnol. J.*, vol. 20, p. 721, 2022, doi: 10.1016/j.csbj.2021.12.032.

[161] D. Xu, S. Zhao, J. Dou, X. Xu, Y. Zhi, and L. Wen, "Engineered endolysin-based 'artilysins' for controlling the gram-negative pathogen Helicobacter pylori," *AMB Express*, vol. 11, p. 63, Apr. 2021, doi: 10.1186/s13568-021-01222-8.

[162] Z. Huang, K. Liu, W. Ma, D. Li, T. Mo, and Q. Liu, “The gut microbiome in human health and disease—Where are we and where are we going? A bibliometric analysis,” *Front. Microbiol.*, vol. 13, p. 1018594, Dec. 2022, doi: 10.3389/fmicb.2022.1018594.

[163] J. N. V. Martinson and S. T. Walk, “Escherichia coli Residency in the Gut of Healthy Human Adults,” *EcoSal Plus*, vol. 9, no. 1, p. 10.1128/ecosalplus.ESP-0003–2020, doi: 10.1128/ecosalplus.esp-0003-2020.

[164] J. Geurtsen, M. de Been, E. Weerdenburg, A. Zomer, A. McNally, and J. Poolman, “Genomics and pathotypes of the many faces of Escherichia coli,” *FEMS Microbiol. Rev.*, vol. 46, no. 6, p. fuac031, Jun. 2022, doi: 10.1093/femsre/fuac031.

[165] M. López-Siles *et al.*, “Prevalence, Abundance, and Virulence of Adherent-Invasive Escherichia coli in Ulcerative Colitis, Colorectal Cancer, and Coeliac Disease,” *Front. Immunol.*, vol. 13, Mar. 2022, doi: 10.3389/fimmu.2022.748839.

[166] M. W. Dougherty and C. Jobin, “Intestinal bacteria and colorectal cancer: etiology and treatment,” *Gut Microbes*, vol. 15, no. 1, p. 2185028, doi: 10.1080/19490976.2023.2185028.

[167] W. T. Cheng, H. K. Kantilal, and F. Davamani, “The Mechanism of Bacteroides fragilis Toxin Contributes to Colon Cancer Formation,” *Malays. J. Med. Sci. MJMS*, vol. 27, no. 4, pp. 9–21, Jul. 2020, doi: 10.21315/mjms2020.27.4.2.

[168] J. Wirbel *et al.*, “Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer,” *Nat. Med.*, vol. 25, no. 4, pp. 679–689, Apr. 2019, doi: 10.1038/s41591-019-0406-6.

[169] L. Sun *et al.*, “Bile salt hydrolase in non-enterotoxigenic Bacteroides potentiates colorectal cancer,” *Nat. Commun.*, vol. 14, no. 1, p. 755, Feb. 2023, doi: 10.1038/s41467-023-36089-9.

[170] L. Lr, S. F, P. V, and G. A, “Commensal Clostridia: leading players in the maintenance of gut homeostasis,” *Gut Pathog.*, vol. 5, no. 1, Aug. 2013, doi: 10.1186/1757-4749-5-23.

[171] P. Guo, K. Zhang, X. Ma, and P. He, “Clostridium species as probiotics: potentials and challenges,” *J. Anim. Sci. Biotechnol.*, vol. 11, no. 1, p. 24, Feb. 2020, doi: 10.1186/s40104-019-0402-1.

[172] S. Jahani-Sherafat *et al.*, “The rate and importance of Clostridium difficile in colorectal cancer patients,” *Gastroenterol. Hepatol. Bed Bench*, vol. 12, no. 4, pp. 358–363, 2019.

[173] D. Ai, H. Pan, X. Li, Y. Gao, G. Liu, and L. C. Xia, “Identifying Gut Microbiota Associated With Colorectal Cancer Using a Zero-Inflated Lognormal Model,” *Front. Microbiol.*, vol. 10, p. 826, Apr. 2019, doi: 10.3389/fmicb.2019.00826.

[174] G. Rogler, “Chronic ulcerative colitis and colorectal cancer,” *Cancer Lett.*, vol. 345, no. 2, pp. 235–241, Apr. 2014, doi: 10.1016/j.canlet.2013.07.032.

[175] H. Lf *et al.*, “Gut Microbiota Signatures in Colorectal Cancer as a Potential Diagnostic Biomarker in the Future: A Systematic Review,” *Int. J. Mol. Sci.*, vol. 25, no. 14, Jul. 2024, doi: 10.3390/ijms25147937.

[176] X. Liu *et al.*, “Blautia—a new functional genus with potential probiotic properties?,” *Gut Microbes*, vol. 13, no. 1, p. 1875796, doi: 10.1080/19490976.2021.1875796.

[177] K. Dubin and E. G. Pamer, “Enterococci and their interactions with the intestinal microbiome,” *Microbiol. Spectr.*, vol. 5, no. 6, p. 10.1128/microbiolspec.BAD-0014–2016, Nov. 2014, doi: 10.1128/microbiolspec.BAD-0014-2016.

[178] A. Davin-Regli, J.-P. Lavigne, and J.-M. Pagès, “Enterobacter spp.: Update on Taxonomy, Clinical Aspects, and Emerging Antimicrobial Resistance,” *Clin. Microbiol. Rev.*, vol. 32, no. 4, Oct. 2019, doi: 10.1128/CMR.00002-19.

[179] J. Zheng *et al.*, “A taxonomic note on the genus Lactobacillus: Description of 23 novel genera, emended description of the genus Lactobacillus Beijerinck 1901, and union of Lactobacillaceae and Leuconostocaceae,” *Int. J. Syst. Evol. Microbiol.*, vol. 70, no. 4, pp. 2782–2858, 2020, doi: 10.1099/ijsem.0.004107.

[180] E. Ghorbani, A. Avan, M. Ryzhikov, G. Ferns, M. Khazaei, and S. Soleimanpour, “Role of lactobacillus strains in the management of colorectal cancer: An overview of recent advances,” *Nutrition*, vol. 103–104, p. 111828, Nov. 2022, doi: 10.1016/j.nut.2022.111828.

[181] A. Muszyński *et al.*, “Granulibacter bethesdensis, a Pathogen from Patients with Chronic Granulomatous Disease, Produces a Penta-Acylated Hypostimulatory Glycero-D-talo-oct-2-ulosonic Acid–Lipid A Glycolipid (Ko-Lipid A),” *Int. J. Mol. Sci.*, vol. 22, no. 7, p. 3303, Mar. 2021, doi: 10.3390/ijms22073303.

[182] A. Sharma, S. Jasrotia, and A. Kumar, “Effects of Chemotherapy on the Immune System: Implications for Cancer Treatment and Patient Outcomes,” *Naunyn. Schmiedebergs Arch. Pharmacol.*, vol. 397, no. 5, pp. 2551–2566, May 2024, doi: 10.1007/s00210-023-02781-2.

[183] C. Herrera-deGuise *et al.*, “Gut Microbiota Composition in Long-Remission Ulcerative Colitis is Close to a Healthy Gut Microbiota,” *Inflamm. Bowel Dis.*, vol. 29, no. 9, pp. 1362–1369, Sep. 2023, doi: 10.1093/ibd/izad058.

[184] S. Camañes-Gonzalvo *et al.*, “Relationship between oral microbiota and colorectal cancer: A systematic review,” *J. Periodontal Res.*, vol. n/a, no. n/a, doi: 10.1111/jre.13289.

[185] D. Yurdakul, A. Yazgan-Karataş, and F. Şahin, “Enterobacter Strains Might Promote Colon Cancer,” *Curr. Microbiol.*, vol. 71, no. 3, pp. 403–411, Sep. 2015, doi: 10.1007/s00284-015-0867-x.

[186] B. W. Wren, “The Yersiniae — a model genus to study the rapid evolution of bacterial pathogens,” *Nat. Rev. Microbiol.*, vol. 1, no. 1, pp. 55–64, Oct. 2003, doi: 10.1038/nrmicro730.

[187] J. K. Triantafillidis, T. Thomaidis, and A. Papalois, “Terminal Ileitis due to Yersinia Infection: An Underdiagnosed Situation,” *BioMed Res. Int.*, vol. 2020, 2020, doi: 10.1155/2020/1240626.

[188] A. I. Akbar, S. George, F. Sobia, and A. Tauseef, “P-032 Yersinia Enterocolitis Mimicking Ulcerative Colitis Flare,” *Inflamm. Bowel Dis.*, vol. 23, no. suppl_1, pp. S15–S16, Feb. 2017, doi: 10.1097/01.MIB.0000512555.35790.71.

[189] Z. Naureen *et al.*, “Bacteriophages presence in nature and their role in the natural selection of bacterial populations,” *Acta Bio Medica Atenei Parm.*, vol. 91, no. Suppl 13, 2020, doi: 10.23750/abm.v91i13-S.10819.

[190] D.-N. Wang *et al.*, “Bacterial infection promotes tumorigenesis of colorectal cancer via regulating CDC42 acetylation,” *PLOS Pathog.*, vol. 19, no. 2, p. e1011189, Feb. 2023, doi: 10.1371/journal.ppat.1011189.

[191] T. Zuo *et al.*, “Gut mucosal virome alterations in ulcerative colitis,” *Gut*, vol. 68, no. 7, pp. 1169–1179, Jul. 2019, doi: 10.1136/gutjnl-2018-318131.

[192] S. X. Ho *et al.*, “Alterations in colorectal cancer virome and its persistence after surgery,” *Sci. Rep.*, vol. 14, no. 1, p. 2819, Feb. 2024, doi: 10.1038/s41598-024-53041-z.

[193] L. Marongiu *et al.*, “Metagenomic analysis of primary colorectal carcinomas and their metastases identifies potential microbial risk factors,” *Mol. Oncol.*, vol. 15, no. 12, pp. 3363–3384, Dec. 2021, doi: 10.1002/1878-0261.13070.

[194] S. Peschel, C. L. Müller, E. von Mutius, A.-L. Boulesteix, and M. Depner, “NetCoMi: network construction and comparison for microbiome data in R,” *Brief. Bioinform.*, vol. 22, no. 4, p. bbaa290, Jul. 2021, doi: 10.1093/bib/bbaa290.

[195] M. L. Tall *et al.*, “*Massilistercora timonensis* gen. nov., sp. nov., a new bacterium isolated from the human microbiota,” *New Microbes New Infect.*, vol. 35, p. 100664, May 2020, doi: 10.1016/j.nmni.2020.100664.

[196] M. R. Bidell, A. L. V. Hobbs, and T. P. Lodise, “Gut microbiome health and dysbiosis: A clinical primer,” *Pharmacotherapy*, vol. 42, no. 11, pp. 849–857, Nov. 2022, doi: 10.1002/phar.2731.

[197] C. C. Kim *et al.*, “Genomic insights from Monoglobus pectinilyticus: a pectin-degrading specialist bacterium in the human colon,” *ISME J.*, vol. 13, no. 6, pp. 1437–1456, Jun. 2019, doi: 10.1038/s41396-019-0363-6.

[198] Y. Wei *et al.*, “Pectin enhances the effect of fecal microbiota transplantation in ulcerative colitis by delaying the loss of diversity of gut flora,” *BMC Microbiol.*, vol. 16, no. 1, p. 255, Nov. 2016, doi: 10.1186/s12866-016-0869-2.

[199] D. Liu *et al.*, “Anaerostipes hadrus, a butyrate-producing bacterium capable of metabolizing 5-fluorouracil,” *mSphere*, vol. 9, no. 4, pp. e00816-23, Mar. 2024, doi: 10.1128/msphere.00816-23.

[200] R. Villemur, M. Lanthier, R. Beaudet, and F. Lépine, “The Desulfitobacterium genus,” *FEMS Microbiol. Rev.*, vol. 30, no. 5, pp. 706–733, 2006, doi: 10.1111/j.1574-6976.2006.00029.x.

[201] L. H. Nguyen *et al.*, “Association Between Sulfur-Metabolizing Bacterial Communities in Stool and Risk of Distal Colorectal Cancer in Men,” *Gastroenterology*, vol. 158, no. 5, pp. 1313–1325, Apr. 2020, doi: 10.1053/j.gastro.2019.12.029.

[202] P. G. Wolf *et al.*, “Diversity and distribution of sulfur metabolic genes in the human gut microbiome and their association with colorectal cancer,” *Microbiome*, vol. 10, no. 1, p. 64, Apr. 2022, doi: 10.1186/s40168-022-01242-x.

[203] “Prevalence of Helicobacter pylori Infection in Colorectal Cancer—a Cross-sectional Study | Indian Journal of Surgery.” Accessed: Oct. 02, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s12262-021-03208-z

[204] P. J. Limburg *et al.*, “Helicobacter Pylori Seropositivity and Colorectal Cancer Risk: A Prospective Study of Male Smokers1,” *Cancer Epidemiol. Biomarkers Prev.*, vol. 11, no. 10, pp. 1095–1099, Oct. 2002.

[205] “Helicobacter pylori promotes colorectal carcinogenesis by deregulating intestinal immunity and inducing a mucus-degrading microbiota signature | Gut.” Accessed: Jul. 10, 2024. [Online]. Available: https://gut.bmj.com/content/72/7/1258.long

[206] T. He, X. Cheng, and C. Xing, “The gut microbial diversity of colon cancer patients and the clinical significance,” *Bioengineered*, vol. 12, no. 1, pp. 7046–7060, doi: 10.1080/21655979.2021.1972077.

[207] X. Kang *et al.*, “Roseburia intestinalis generated butyrate boosts anti-PD-1 efficacy in colorectal cancer by activating cytotoxic CD8+ T cells,” *Gut*, vol. 72, no. 11, pp. 2112–2122, Nov. 2023, doi: 10.1136/gutjnl-2023-330291.

[208] K. Nie *et al.*, “Roseburia intestinalis: A Beneficial Gut Organism From the Discoveries in Genus and Species,” *Front. Cell. Infect. Microbiol.*, vol. 11, p. 757718, 2021, doi: 10.3389/fcimb.2021.757718.

[209] M. I. Moreira de Gouveia, A. Bernalier-Donadille, and G. Jubelin, “Enterobacteriaceae in the Human Gut: Dynamics and Ecological Roles in Health and Disease,” *Biology*, vol. 13, no. 3, Art. no. 3, Mar. 2024, doi: 10.3390/biology13030142.

[210] N. O. Kaakoush, “Insights into the Role of Erysipelotrichaceae in the Human Host,” *Front. Cell. Infect. Microbiol.*, vol. 5, Nov. 2015, doi: 10.3389/fcimb.2015.00084.

[211] G. Liu *et al.*, “Improved diagnostic efficiency of CRC subgroups revealed using machine learning based on intestinal microbes,” *BMC Gastroenterol.*, vol. 24, no. 1, p. 315, Sep. 2024, doi: 10.1186/s12876-024-03408-3.

[212] Y. Ulger, A. Delik, and H. Akkız, “Gut Microbiome and colorectal cancer: discovery of bacterial changes with metagenomics application in Turkısh population,” *Genes Genomics*, vol. 46, no. 9, pp. 1059–1070, Sep. 2024, doi: 10.1007/s13258-024-01538-2.

[213] R. Ichimura, K. Tanaka, G. Nakato, S. Fukuda, and K. Arakawa, “Complete genome sequence of Mediterraneibacter gnavus strain RI1, isolated from human feces,” *Microbiol. Resour. Announc.*, vol. 0, no. 0, pp. e00863-24, Sep. 2024, doi: 10.1128/mra.00863-24.

[214] M. A. Morgan and E. J. Goldstein, “Bulleidia extructa: An underappreciated anaerobic pathogen,” *Anaerobe*, vol. 69, p. 102339, Jun. 2021, doi: 10.1016/j.anaerobe.2021.102339.

[215] S. J. Quillin and H. S. Seifert, “Neisseria gonorrhoeae host-adaptation and pathogenesis,” *Nat. Rev. Microbiol.*, vol. 16, no. 4, pp. 226–240, Apr. 2018, doi: 10.1038/nrmicro.2017.169.

[216] N. J. Weyand, “Neisseria models of infection and persistence in the upper respiratory tract,” *Pathog. Dis.*, vol. 75, no. 3, p. ftx031, Apr. 2017, doi: 10.1093/femspd/ftx031.

[217] Z. Hexun *et al.*, “High abundance of Lachnospiraceae in the human gut microbiome is related to high immunoscores in advanced colorectal cancer,” *Cancer Immunol. Immunother.*, vol. 72, no. 2, pp. 315–326, Feb. 2023, doi: 10.1007/s00262-022-03256-8.