On
uncovering and understanding
interactions between
bacteriophages and bacteria

Members of the dissertation committee:

Dr. H. Herrema, *Amsterdam Universitair Medische Centra*
Prof. dr. C. Hill, *University College Cork*
Prof. dr. M.A. Huijnen, *Radboud Universiteit Nijmegen*
Prof. dr. R.J.L Willems, *Universiteit Utrecht*
Prof. dr. A. Zhernakova, *Rijksuniversiteit Groningen*

Paranymphs:

J.N.A. Vink
R.E. McKenzie

The research in this thesis was funded by the Dutch Research Council (NWO) under project number 864.14.004.

ISBN: 978-90-830912-2-8

Cover: Patrick A. de Jonge & Print Service Ede - Ede
Printing: Print Service Ede - Ede

On
uncovering and understanding
interactions between
bacteriophages and bacteria

Over het ontdekken en begrijpen van interacties
tussen bacteriofagen en bacteriën
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 26 oktober 2020 des ochtends te 11.00 uur

door

PATRICK ALEXANDER DE JONGE

geboren op 15 juli 1989
te Arnhem

Promotor:

Prof. dr. B. Snel

Copromotoren:

Dr. B.E. Dutilh
Dr. S.J.J. Brouns

Voor en door Jan en Marty

Table of contents

1

General introduction and thesis outline

"Go not to the Elves for counsel, for they will say both no and yes."

The Fellowship of the Ring; J.R.R. Tolkien

Bacteriophages – the most abundant biological entities

What are bacteriophages?

Bacteriophages (or phages) are viruses. Specifically, they are viruses that infect bacteria (from Greek *phagos* meaning devouring, so literally "bacterium devourer"). Being viruses, phages are neither alive nor dead, although how much of either is hotly debated[1–3]. Viable answers to the question "are viruses/phages alive" include "yes" (e.g. [3,4]) and "no" (e.g. [1]). Arguably, because the answer to this question depends on your definition of "life", the question itself is too irrelevant to answer[5]. In the end, it is safest to state that viruses (and by extension phages) are a form of biological matter that are *different* from cellular life (e.g. humans, bacteria, and jellyfish). Cellular life forms by definition have lipid membrane(s) surrounding an instruction manual (DNA). They additionally have tools to copy them, keep them intact, sense their surroundings, and make energy. Viruses, meanwhile, only have the instruction manual, something surrounding it, and something to grab a host with. They cannot copy or maintain themselves but are inert particles that cannot exist without their hosts.

What do bacteriophages look like?

This depends on the species of bacteriophage. Like all viruses, the basis of a phage is its instruction manual or genome. This genome is most often (to current knowledge, see "Viral dark matter of the biosphere") DNA. Some known phages have RNA, a molecule similar to DNA. Phage genomes are small compared to cellular genomes. The genome of phage lambda (λ, 48,502 base pairs) is for example a thousand times smaller than that of model bacterium *Escherichia coli* (~4.2 million bp), and about one hundred thousand times smaller than the human genome (~3.3 billion bp). Of course, this is just one phage species. But while phage genome sizes have a huge range, from 3,569 bp[6] to over 700,000 bp[7,8], even the largest ones are small compared to most cellular life.

As phage genomes would quickly break down if left exposed for too long, they are protected in a protein shell or capsid. This capsid is generally a 20-sided shape called an icosahedron (Figure 1), which has the best surface area-to-volume ratio of shapes made out of multiple identical flat surfaces (a Platonic solid)[9]. Capsids of most phages are decorated with antenna-like molecules that often help in binding to hosts[10] or other structures[11]. Decorated icosahedral capsids form the whole particle of some smaller phage species, like *Microviridae.* Many others have larger tail structures that are used for binding. Such tailed phages form the

Caudovirales (*cauda* is Latin for tail), which is currently the largest taxonomic group of phages. *Caudovirales* tails come in three shapes (Figure 1)[12]. The first of these three look like long, flexible needles and phages that have them form the *Siphoviridae* family (*sipho* is Latin for needle). Next are those that only contain the bottom section used for binding and as a result are very short. These are carried by phages in the *Podoviridae* family (*podi* is Greek for foot). Finally, there are the most complex tails, carried by phages in the *Myoviridae, Herelleviridae* and *Ackermannviridae* families. Their tail is composed of a rigid inner tube and an outer shell that works like a coil spring under tension. This tension is released upon infection, rapidly contracting the outer shell so that the inner tube punches a hole into the host. This contraction is the origin of the *Myoviridae* name (*mys* is Greek for muscle). Nothing in biology is absolute (except this statement), and it follows that not all phages have icosahedral heads. Some have entirely different shapes. These are the filamentous and pleomorphic phages[13]. The former unsurprisingly are shaped like filaments, while the latter are the "others" category of phage shapes[13]. Due to this diversity in particle shape, phage particle sizes have a large range, from small icosahedrons of 20 nanometre in diameter to tailed particles of 150 by 180 nm[14].

How do bacteriophages work?

In essence, all bacteriophages introduce their genome into a host cell and then produce offspring. They do this through several different life cycles. Phage life cycles include two well studied and two less well studied varieties. In every variety, however, phage life cycles are composed of four basic stages. These are 1) meeting a host and infecting it with the phage genome, 2) evading the defences of the host, 3) reproduction, 4) releasing new phages into the environment. In the well-studied lytic and lysogenic life cycles, this basic four-stage process is shaped differently. In both cycles, the phage attaches to the host, injects its genome into it, and counters or evades any host mechanisms that try to destroy the genome. Here the two life cycles diverge. In the lytic cycle (done by lytic phages), new phage particles are made from host cell material. At the same time proteins that can make holes in the cell wall are produced. Activation of these proteins, which can happen in different ways[15], starts the final stage. The cell wall is ruptured, and new phage particles released into the environment.

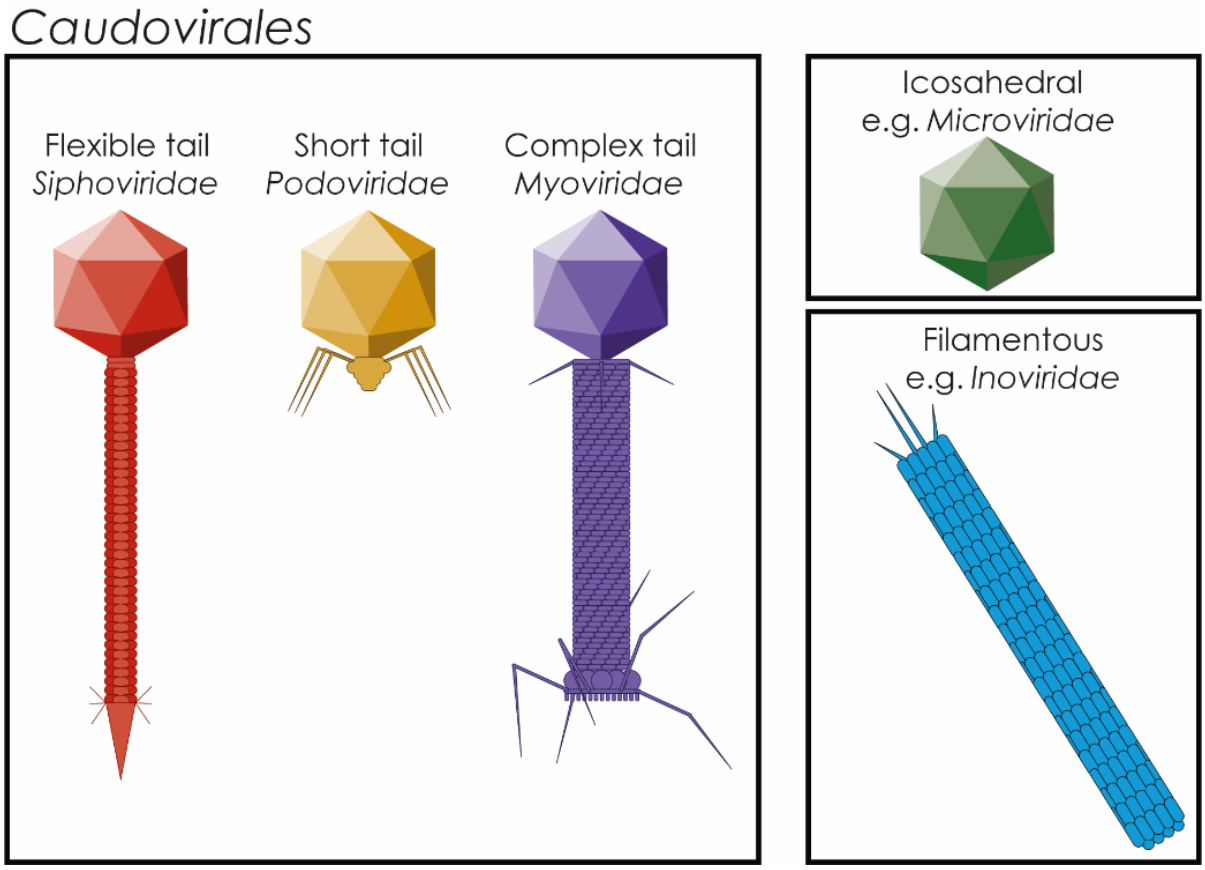


Figure 1: The main morphologies of bacteriophages. *The Caudovirales are tailed phages with either flexible tails (Siphoviridae such as lambda), short tails (Podoviridae such as P22), complex tails (Myoviridae such at T4). The main untailed phage morphologies are icosahedrons (such as Microviridae phage phiX174) and filamentous phages (such as Inoviridae phage M13).*

The main difference between the lytic and lysogenic life cycle occurs after genome entry into the cell. At this point of the lysogenic cycle (done by temperate phages), the phage genome integrates into the host genome, at which point it is called a prophage. By forming a prophage, a temperate phage co-exists with its host without harming it. A bacterium with a prophage is a lysogen. The advantage to existing as a prophage is that you lift along with your host, and freely spread through a population when your host divides in two. It can also be dangerous, however, as random mutations can corrupt the prophage genes that encode instructions for turning back into a full phage particle. This can turn a prophage into a permanent part of the host genome[16]. Assuming that the prophage remains intact, certain signals can trigger its excision from the host genome. Such signals are often stressors to the bacterium, such as DNA damage by UV radiation, environmental changes (e.g. temperature, salinity, or starvation), and certain pollutants[17,18]. Not all excision signals are environmental stressors. Some prophages cannot

make capsids, and instead excise when a second phage produces particles and hijacks these particles[19]. Except for these instances, excised prophages proceed in identical fashion to the lytic cycle. They produce new phage particles from host cell material and break out of it. Many phages are both lytic and temperate, and "decide" which life cycle to pursue upon infection[20,21].

The two less well studied life cycles are chronic and pseudolysogenic. Filamentous phages engage in chronic infections by continuously producing new phages that are released from their host cells without killing them. Pseudolysogeny can be thought of as phages being on stand-by. Instead of incorporating into the host genome or dividing, the phage genome is dormant some period of time before proceeding into lysogenic or lytic life cycles[22]. As chronic and pseudolysogenic infections in and of themselves do not result in lysing or lysogenising cells, they are difficult to study. As a result, their impact in nature is mostly unclear.

Where are bacteriophages?

Bacteriophages are everywhere on the planet where bacteria are found. By extension this means that they are everywhere on the planet where there is life. Right now, assuming you are a healthy human adult, there are about 2 trillion phages just in your intestinal tract[23]. There are also phage populations of unknown size in or on your skin[24], lungs[25,26], mouth[27] and urinary tract[28], among others. Outside of the human body, there are phages even in harsh environments like salty lakes[29], hot water vents on the ocean floor[30], and Antarctic lakes under 2-6 meters of ice[31].

Depending on the life cycle used, phages can either actively spread offspring (lytic and chronic infections), or not (lysogenic and pseudolysogenic infections). This means that besides free phage particles being everywhere where there is life, phages and phage fragments are also inside bacteria. To be precise, this includes prophages, pseudolysogenic phages, and actively replicating phage particles (e.g. filamentous and lytic phages). The fraction of bacteria that carry phage genomes, particularly prophages, at any given time remains unclear. It is certain that this greatly depends on the environment[18,32,33]. This is exemplified in the human intestinal tract. A recent model suggests that deep in the mucus layers covering the intestinal walls bacterial densities are low and lysogen formation is rare, so few bacteria have prophages[34]. Meanwhile, the opposite is true higher up in the mucus layers. Here and in general, it is also thought that the fraction of bacteria carrying prophages is related to bacterial density. The existence of such a link is uncertain, however, as thorough re-analysis of earlier data recently showed[35]. In fact, there are question marks attached to all models that seek to explain which factors regulate prophage formation rates. This is not helped by the estimates of prophage rates in nature ranging

from 0% to 100%[36]. These numbers vary so widely because they depend on the environment from which the sample is taken, the time at which it is taken, and the methods used to analyse it. All this being said, recent research seems to agree that a sizeable fraction of bacteria in a variety of ecosystems contain prophages[18,32,33].

How many bacteriophages are there?

There are a lot of bacteriophages. To understand how many, first a reminder on phages and viruses. As stated above (see "what are bacteriophages?") phages are a subdivision of viruses. When talking about the number of phages on Earth, we often talk about the number of viruses instead (e.g. [36,37]). This is because we assume most viruses on Earth are phages, because bacteria are the most common cellular life form on Earth[36]. In the next paragraph referrals of viral counts are thus mostly equal to phage counts.

For the longest time, up to about the late 1980s, we thought viruses were rather rare[38]. Then, a Scandinavian research group fluorescently labelled viruses and counted them under a microscope[39]. While acknowledging that not every particle that they saw was necessarily a virus (which resulted in the common term virus-like particles or VLPs), they concluded that viruses were much more abundant than previously thought[39]. In fact, thirty years on from this initial insight, we know that viruses are not rare at all. They are instead the most abundant form of biological matter on the planet, with a total of 4.8×10^{31} particles[36]. This number is essentially unimaginably big. In an attempt to illustrate the size of 4.8×10^{31}, compare it to the estimated number of stars in the known universe. High estimates of this are at about 10^{24} stars[40]. This means that there are 10 million viruses on planet Earth for every star in the known universe. While this is admittedly still somewhat abstract, the core facts are that viruses are more common than any life form and most viruses are phages.

There are three notable caveats to the above estimate of viral abundance. The first is that it is an estimate based on VLP-to-bacteria ratios in various environments[36]. While certain environments are well characterised, most notably the seas and oceans[41], others are much less so. This includes estimates of phages in terrestrial samples, from which it is technically challenging to accurately extract all VLPs[36]. The second caveat is that the estimate refers to VLPs. As stated, not all VLPs are phages or even viruses. Some VLPs are instead particles that look like them, like pyocins which are entities produced by bacteria and used to kill other bacteria[42]. Finally, the third caveat is that this VLP estimate only concerns free phage particles. However, as described in the previous section ("Where are bacteriophages"), a sizeable

proportion of bacteria carry prophages or other phage particles. These phages are not counted as free particles, and adding them to the current estimate of viral abundance could increase it sizeably[36].

Impacts of phages

Now that we know that phages are the most abundant biological entity in the biosphere, the obvious next question is: What do all these phages do to environments? For convenience, I will answer this question in two parts, first focussing on general ecology, and then human health.

What do bacteriophages do?

Phages impact every microbial ecosystem. They do this in two main ways: altering bacteria and killing bacteria.

Phages alter bacteria in several ways. First, processes inside bacteria are changed during lytic phage infections. As some lytic phage infections can last for hours[43], this can have a large effect on bacterial communities. For example, in photosynthetic bacteria a phage infection can hijack photosynthesis and divert the energy to phage particle production[43]. The second form of alterations are those that result from phage genome integrations. Prophages sometimes carry genes that add new functions to bacteria. This includes photosynthesis genes that improve bacterial energy production[44,45]. Prophages can also carry genes that protect the host cell from infection by other phages (superinfection exclusion)[46,47]. Traits obtained through prophages generally help a bacterium compete with others[44]. This is beneficial to the phage, as their survival is linked to that of the hosts. However, what helps the host bacterium can also help make humans ill, by for example introducing toxins[47,48]. As a third way that phages alter bacteria, they can transfer genes between two different bacteria. This is called horizontal gene transfer, as opposed to vertical gene transfer by making offspring. Phages transfer genes horizontally through generalised and specialised transduction. In generalised transduction phage particles are erroneously packed by host DNA[49]. These particles then infect other bacteria, transferring DNA from one host to another. A generalised transducing phage sometimes loses the ability to pack their own DNA, thereby becoming a gene transfer agent[50]. Specialised transduction occurs when excising prophages also cut out DNA adjacent to them on the bacterial genome, after which it is packaged into phage capsids[51]. Through all of the

above mechanisms phages alter the dynamics of bacterial communities and are key players in bacterial evolution[50].

In addition to altering bacteria, phage-mediated killing of bacteria has major impacts on microbial communities. Phage-mediated killing of bacteria is common: an estimated 20% of marine bacteria are killed by phages every day[52]. Various studies suggest that 9% to over 50% of bacteria in soil and marine environments are infected by a phage at any given time[53–55]. After a bacterium is killed by a phage, other bacteria can either take the place (i.e. ecological niche) of the killed bacterium, or they can feed on the debris of the lysed bacterium. Through both mechanisms, phages serve as controlling agents in microbial communities, which has effects on a grand ecological scale. Globally, lysis by phages releases about 1 billion tonnes of carbon per year[56,57]. In marine environments the nutrients released by viral lysis sinks into the deeper oceanic layers, providing a major carbon source to the ecosystems of the deep ocean[56]. A similar process of freeing nutrients through phage lysis likely makes it possible for plants to compete with bacteria over nitrogen sources[58]. This process is called the viral shunt, and together with horizontal gene transfer is a major driver in maintaining microbial species diversity[52,59]. Through the viral shunt, phages have a major impact on the global cycles of elements that are essential for life, such as carbon and organic nitrogen.

What do bacteriophages do to humans?

Many indirect health-related things, although what is included in this is currently a hot topic of study. Humans, quite logically, tend to be most interested in things that affect humans. Strangely, there is arguably more knowledge about phages from the middle of the Pacific Ocean[41] than about those that live in our own gut[60]. This, in my opinion, has two reasons.

Firstly, phages do not directly impact our health in the way that other viruses (e.g. influenza viruses like the Spanish Flu and SARS-coronaviruses) or bacteria (e.g. *Yersinia pestis* a.k.a. the Black Death) do. Instead, phages indirectly affect us by interacting with the bacteria that live in and on us. For example, the bacteria in your gut directly aid digestion by breaking down complex molecules from food into simpler compounds that your body can use to generate energy[61]. Remove these bacteria, and you would not be able to digest the food you eat. Meanwhile, phages regulate those bacterial populations through the mechanisms explained in the previous section (see “What do phages do?”). This means that 1) the importance of phages to human health was recognised only recently because it is diffuse and indirect, and 2) the precise effects are difficult to study. This is exemplified in studies of crAss-like phages. This

phage family is the most common group of phages in humans, found in an estimated 50% of individuals[62]. However, despite intensive studies, their impact on gut environments and human health remains a mystery.

Human-associated phages are also understudied because of technical difficulties. Phage samples from inside the human body are hard to ethically obtain from healthy individuals. Most studies on human gut phage communities therefore sample faecal matter. However, the intestinal tract is long and has great variety in aeration, nutrients, and other factors along its length[63]. In most of the intestinal tract there is furthermore no oxygen, and to bacteria that live in it oxygen is often toxic[64]. This makes them challenging to isolate and cultivate in a laboratory setting. In turn, phages that infect these bacteria are rarely isolated, and therefore largely remain unknown.

While human gut phage environments are indeed understudied, recent years have seen a rise in interest. This has shown that gut phage populations are specific to individuals, although some core members are commonly found (e.g. crAss-like phages)[65]. Furthermore, research is uncovering the composition and dynamics of gut phage communities, where different phage survival strategies seem to be applicable to different parts of the intestine[34,60]. This attention is also providing the first implications that gut phage communities are involved in inflammatory bowel diseases[66–68], with malnutrition[69], and can be altered by dietary components[70]. It is yet still very difficult to interpret such results. In the case of inflammatory bowel disease, for example, it is unclear whether alterations in phage communities are caused by the disease or vice versa. Thus, much additional research is still needed, not least on phage characterisation.

Viral dark matter

The previous sections provided a very general overview of the past century in phage biology, showing the sizeable extent of our knowledge. While we evidently know much, all this knowledge has been gained from only a sliver of the existent phage biodiversity. The grand majority of phages remains unstudied, in what has dramatically been dubbed viral dark matter[42,71]. To illustrate this point, we will first look at phages that have already been characterised.

Characterised phages

The International Committee on the Taxonomy of Viruses (ICTV) serves as an authority on what is a viral species[72]. As of early 2020, there were 5,560 entries on their list of viral species. Although most viruses on Earth are phages (see "How many phages are there"), only 1,394 (25%) of the ICTV-recognised viral species were bacteriophages. This imbalance is largely because humans focus on viruses that directly harm us or our societies by killing us, our food crops, or our cattle.

Within the collection of ICTV-accepted phage species there are further biases. One such bias is in the material making up phage genomes. Phage genomes are made from either DNA or RNA. These molecules are made from either one strand of genomic code (single-stranded or ss) or two matching strands (double stranded or ds). Among the 1,413 phage species (Figure 2A), 1,329 carry dsDNA (95.3%), 54 carry ssDNA (3.9%), 7 carry dsRNA (0.5%), and 4 carry ssRNA (0.3%). The dsDNA phage species are further skewed toward the tailed *Caudovirales* (see "What are phages"), which alone account for 1,320 or 94.7% of the ICTV-recognised phage species.

Next to this bias toward *Caudovirales*, characterised phage species are also biased toward only a few bacterial hosts (Figure 2B). Five bacterial genera (*Mycobacterium, Escherichia, Pseudomonas, Salmonella,* and *Bacillus*) serve as hosts for 647 or 46.4% of the ICTV-recognised phages, whereas the remaining 747 phage species infect 97 bacterial genera. Not incidentally, these five genera all contain species that are easy to work with. Some of them are also important human pathogens (e.g. *Mycobacteria* cause tuberculosis and *Salmonella* cause typhoid fever). Meanwhile, no ICTV-recognised phages infect *Bacteroides*, which dominate microbial human gut communities[64]. At a higher taxonomic level, the phylum *Bacteroidetes* are infected by 24 species (1.7%) in the ICTV list. This discrepancy is caused by the difficulty in working with *Bacteroidetes*, to which oxygen is toxic, and the fact that they generally do not cause disease. In summary, characterised phage collections reflect the not unreasonable tendency to focus on phages that infect bacteria that are easy to work with and that cause problems to humans. It does not reflect true natural phage biodiversity.

Likewise, the bias toward *Caudovirales* does not automatically reflect true biodiversity. Instead, it probably reflects that most techniques that we use to study phages select for *Caudovirales* (see "The problem of who is who" and **Chapter 3**). In fact, purposeful efforts uncovered a wealth of RNA phages[73], ssDNA phages[74,75], non-tailed dsDNA phages[76], and filamentous ssDNA phages[77]. It is these phages (and, admittedly, probably a lot of undescribed

Caudovirales members) that make up viral dark matter. This is illustrated by the fact that the ICTV-recognised species represent less than 0.00002% to the 100 million phage species estimated to exist on planet Earth[36]. There are also other databases of phages that are not (yet) officially recognised. The National Centre for Biotechnology Information GenBank database, is such a collection. While GenBank sequences likely include sizeable redundancy, it contains several tens of thousands of phage genomes. But while about ten times larger than the ICTV lists, this still means that about 0.0002% of phage species are currently described. By whatever measure that is used, characterised phage biodiversity forms a tiny fraction of what exists in the biosphere. Hence, a major focus of modern phage biology is to better characterise viral dark matter.

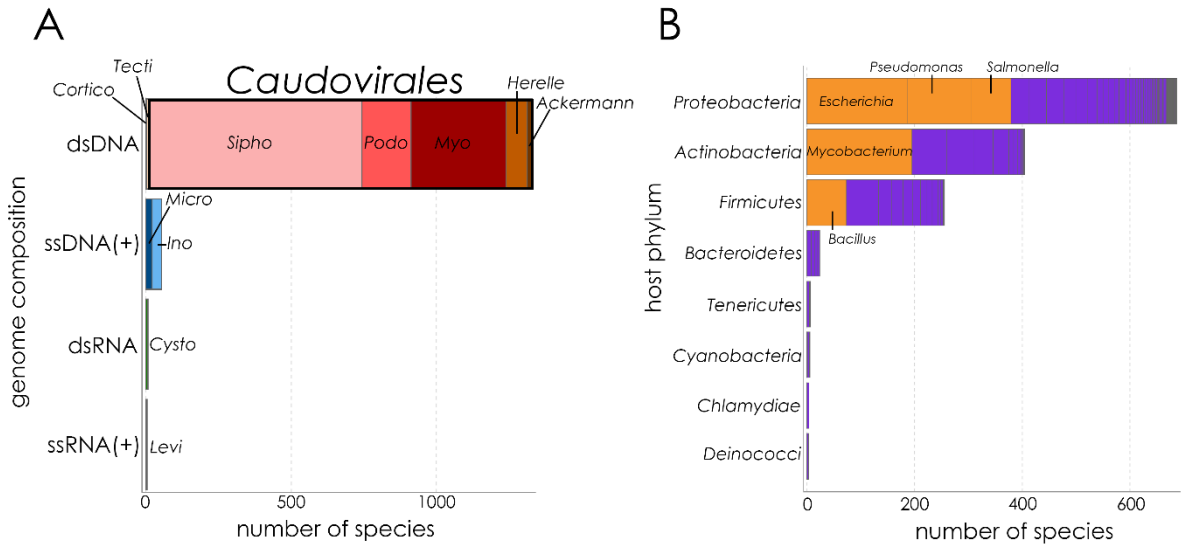


Figure 2: Biases in International Committee on the Taxonomy of Viruses (ICTV)-recognised phage species. (A) *ICTV-recognised species are predominantly tailed Caudovirales. Genome composition denotes the type of genetic material used by the phage: double stranded DNA (dsDNA), double stranded RNA (dsRNA), single stranded DNA (ssDNA) and single stranded RNA (ssRNA). Small type italic labels are taxonomic family name prefixes, which are followed by the suffix -viridae.* ***(B)*** *Most ICTV-recognised species infect within the phyla Proteobacteria, Actinobacteria, and Firmicutes. Especially overrepresented are hosts in the orange-coloured genera*: *Escherichia, Pseudomonas, Salmonella, Mycobacterium, and Bacillus. Data according to the ICTV master species list, retrieved in February 2020.*

Analysing viral dark matter through metagenomics

The best tool currently available for uncovering viral dark matter is metagenomics, which is the study of all genomes in a community (a metagenome). Viral metagenomics developed over the last two decades with plummeting costs of genome sequencing (i.e. determining DNA sequences). Using metagenomics approaches, phages can now be studied without isolating them in a laboratory setting. This is advantageous because most bacteria have not been grown in laboratory settings[78,79], which limits phage isolation efforts. Besides culture-independent studies, viral metagenomics allows for any phage community to be fully sequenced. This provides huge amounts of data. Hence the difference between the thousands of phage species recognized by the ICTV and the tens of thousands of phage genomes in GenBank. The former are highly curated and often well-studied phages, while the latter are mostly phage genomes that are less well studied and curated. In addition to GenBank there are large and continuously growing databases consisting of metagenomes from whole viral communities[80]. These databases cause the central conundrum of modern phage biology: the ability to obtain metagenomic data is virtually unconstrained, but the ability to understand the underlying communities is very much constrained. Among the insights that are currently lacking because of this are the inability to rapidly identify new phages and the inability to link phages to their hosts.

The problem of who is who

The first step to characterising (part of) an environmental phage sequence is to compare it to already characterised phages. But only a sliver of phages have been studied at all, let alone in detail. As a result, new phage sequences are often unlike anything in databases of characterised phages. They might be similar to other uncharacterised phages, but this often does not provide new information. Thus, a priority in modern phage biology is to expand our databases of characterised sequences.

Classic methods to discover and characterise phages depend on several tried-and-true methods, foremost of which is the plaque assay[81]. In a plaque assay, a target bacterium and a (mix of) phage(es) are mixed into agar solutions and spread on a petri dish. The agar contains all the nutrition the bacterium needs and while solid, has a density that is low enough to allow both bacterium and phage to diffuse through it. The result is growth of an opaque lawn of bacteria, except for clear spots (plaques) where phages have infected and killed bacteria (for an example, see Figure 3a in **chapter 5**). From such plaques, phages can be retrieved and

transferred to a fresh petri dish. Multiple rounds of this results in phage isolation. While plaque assays are simple, they have important drawbacks that cause the skewed nature of characterised phages (see "Characterised phages"). These include a dependency on bacteria that are cultivable and easily form lawns in petri dishes, and on phages that form visible plaques[81]. This already leaves out most bacteria and all non-lytic phages. Additionally, these techniques select for the phages that infect fastest and produce the most offspring (see **chapter 3**). Finally, every single phage-host combination needs a separate assay when using such techniques, making them highly laborious.

After phage isolation follows the next challenge: understanding how it works. Phage genomes, like all genomes, consist of genes that encode proteins. Each protein has a function, like capsid proteins forming the phage head, polymerases making DNA, or lysins cutting holes in the cell envelope at the end of infection. Like phage genomes, the first approach to understand the function encoded by a gene is to compare it to what is already known. Here too the prevailing problem is the skewed nature of characterised phage collections. Proteins from phages that infect *Proteobacteria* are generally more similar to each other than for example to proteins from phages that infect *Bacteroidetes*. This is further compounded by phage evolution, which is thought to be rapid and involves much horizontal gene transfer. Comparing proteins from a non-isolated phage to the existing database therefore often results in limited new information. As a result most phage proteins remain without a known function[82]. One solution is to isolate phage proteins and test their function one by one, but this is not possible for all proteins and even when possible is slow and laborious. In addition, computational predictive techniques of protein functions are not yet capable to accurately predict function from genome sequences alone.

Due to the above drawbacks in phage isolation and study, recent research has sought to develop alternative methods that incorporate metagenomics. These new methods range from purely computational to using extensive "wet-lab" (i.e. non-computational) approaches. An example of how such method can lead to new insight can be derived from studies of crAss-like phages (see "what do bacteriophages do to humans"). The original representative of this widespread phage family was crAssphage, which was discovered by an in-depth computational analysis of an existing dataset that was composed of gut viral communities of four sets of identical twins and their mothers[83,84]. Subsequent computational data-mining approaches uncovered a wide variety of related phages composing the new crAss-like phage family[85,86]. Sophisticated wet-lab enrichments then led to the isolation of the first crAss-like phage[87], while mixed computational/wet-lab approaches revealed the world-wide distribution of crAss-like

phages[62]. The above, while impressive, still involved half a decade of labour by several research groups. Development of new techniques to rapidly characterise new phage species would therefore be highly useful. Development of such techniques, besides characterising phages, often also focus on a second insight lacking in modern phage biology, which I will cover next.

Interactions between phages and hosts

Next to phage identification, another crucial insight that is lacking is understanding from metagenomic data sets which phage infects what bacterial species. In turn, this is dependent on understanding the mechanisms used by phages to interact with their hosts.

Phage meets host, phage likes host

As stated at the start of this piece, phages are pieces of genetic material, protein, and sometimes lipid floating around. Floating around here refers to Brownian motion, which are movements caused by heat and particles hitting each other. As this movement is random, the process by which phages and hosts meet is also random. Even gut phages that weakly bind to mucus to increase chances of meeting bacteria that eat said mucus[11], merely randomly move through the mucus at a slower pace. By chance, this random motion can cause a phage to bump into objects. Such objects are sometimes host bacteria. More precisely, such object are sometimes certain structures on bacterial surfaces: receptors.

Different molecules can be receptors for phages[12]. They can be proteins on the cell surface, pili (hair-like appendages), flagella (whips used to move), lipopolysaccharide chains (appendages often carrying toxins), and teichoic acids (structural elements). Almost everything on the bacterial surface is used by some phage or another as a receptor. But although everything serves as a receptor for *a* phage, single phage species cannot use everything as a receptor. This is a central facet of phage biology: known phages are almost exclusively specific to one receptor. Although other factors also play a role here (see **chapter 3**), this generally means most phages infect one host species. To this central facet there are multiple caveats, like 1) a minority of characterised phages are not specific at all, and 2) it is based on the tiny fraction of phage species that have been studied. While we are not sure about the specificity among viral dark matter, there are multiple arguments for why these phages are likely mostly specific, and multiple for why many might not be specific at all (see **chapter 3**). But based on current knowledge it seems that phage specificity to a receptor and host is the norm[88]. Because of this

specificity, the phage-receptor interaction is useful to a wide variety of phage studies (see **chapters 4 and 5**).

If a phage bumps into its receptor, it might bind to it. This binding to receptors is done by the aptly named receptor-binding proteins, which are located either on the capsid (in e.g. *Microviridae*) or at the bottom of the tail (in *Caudovirales*). What follows differs among phage species, but it often involves a two-staged process[12]. Stage one is the random walk or irreversible binding. The receptor-binding proteins attach to and detach from receptors at random, which can cause the phage to "walk" along the surface[89]. This can have two outcomes. Either all the receptor binding proteins all release from receptors and the phage floats away, or the second binding stage is triggered. What this trigger is again depends on the phage species. It might be the simultaneous binding of multiple receptor binding proteins (e.g. in phage T4[12]) or a rotational movement of the phage particle (e.g. in phage lambda[90]). Whatever the trigger is, when it happens phage binding becomes irreversible. Once irreversible, physical changes in the particle culminate in the phage injecting its DNA into the host. In *Myoviridae*, these changes cause the tail to collapse and punch a hole into the cell envelope, through which the DNA is pushed by the high pressure inside the phage head[91]. Other phages inject their DNA through pre-existing channels in the cell wall or otherwise dig a hole into the cell[12]. What follows are the stages of infection mentioned earlier (see "How do phages work").

Predicting hosts

As explained above, a phage is a virus and therefore dependent on its bacterial host for existence. Thus, the study of a phage is incomplete without host identification. Of course, the best method to uncover phage-host pairs is to isolate phages under laboratory conditions using plaque assays. But as explained above, this has severe limitations. Therefore, multiple computational methods to predict which phage infects which host have been developed in the past decade. These often look at traces of past interactions that are left in the phage or host genome. Such traces can be traits shared by phage and host such as similar abundances across environments, general patterns in the DNA code (tetranucleotide frequency), shared whole genes, and direct similarity to prophages[92]. Another computational approach is to exploit a bacterial defence mechanism called CRISPR-Cas, which maintains a database of DNA pieces from entities that previously invaded a bacterium[93]. While there are many computational host-prediction methods, many share a single drawback: lack of reliability[92]. The exception to this is CRISPR-Cas, which evolved to recognise specific phages (among others) encountered by a

host. However, most bacteria do not have CRISPR-Cas systems[94], and the ones that do often link nowhere due to rapid phage evolution. Thus, at least currently, accurate phage-host prediction needs a mixture of computational and wet-lab methods. Development of such methods was the main goal of the research presented in this thesis.

Outline of the thesis

The central focus of the research presented in this thesis is to understand and uncover the interactions between bacteriophages and bacteria. The research presented in the following four chapters looks at different aspects of phage biology and presents novel methods to study and uncover phage-host interactions.

Chapter 2 details research into *Bacteroides associated carbohydrate-binding often N-terminal* (BACON) protein domains in crAss-like phages. In a probable tail protein, crAssphage carries a repeat of putative carbohydrate-binding BACON domains that may be associated to binding the crAssphage host. We use these BACON domains to study the biodiversity of crAss-like phages and identify several novel gut phage lineages. Additionally, we provide a comprehensive study on BACON-domain evolution in crAss-like phages.

Chapter 3 contains a literature review on bacteriophage host-range. After providing a framework to understand phage host range, this chapter describes the molecular mechanisms that phages have evolved to allow them to infect multiple bacterial hosts. In addition, the methodological biases that may cause misinterpretation of phage-host interactions are described.

Chapter 4 presents research on the development of a novel methodology for high-throughput studies of phage-host interactions from metagenomic data. In it, we exploit the fact that phage-host interactions are dependent on phages binding to receptor molecules on the host cell surface. By isolating host cell membranes and exposing them to environmental phage mixtures, we separate phages based on their ability to bind a host species. We show that this method is viable using model phages, and subsequently apply the method to uncover phage-host interactions in an environmental sample.

Chapter 5 portrays efforts to exploit styrene maleic acid lipid particles (SMALPs) as a tool to study phage-host interactions. SMALPs are a novel method that allow the isolation of bacterial membrane proteins while maintaining them in their native membrane environments. We show that caught in SMALPs, membrane receptors still inactivate multiple model phages. Furthermore, we show that these interactions result in phage genome ejection. We conclude that SMALPs could see future use for detailed molecular studies of phage-host interactions.

Chapter 6 offers concluding remarks on the work presented in this thesis and offers outlooks on the field of phage biology.

2

Evolution of BACON domain tandem repeats in crAssphage and novel gut bacteriophage lineages

"Deeds will not be less valiant because they are unpraised."

The Return of the King; J.R.R. Tolkien

Published as: de Jonge, P.A.; von Meijenfeldt, F.A.B.; van Rooijen, L.E.; Brouns, S.J.J.; Dutilh, B.E.; *Evolution of BACON Domain Tandem Repeats in crAssphage and Novel Gut Bacteriophage Lineages*; Viruses; 2019; 11; 1085.

Abstract

The human gut contains an expanse of largely unstudied bacteriophages. Among the most common are crAss-like phages, which were predicted to infect *Bacteriodetes* hosts. CrAssphage, the first crAss-like phage to be discovered, contains a protein encoding a *Bacteroides*-associated carbohydrate-binding often N-terminal (BACON) domain tandem repeat. Because protein domain tandem repeats are often hotspots of evolution, BACON domains may provide insight into the evolution of crAss-like phages. Here, we studied the biodiversity and evolution of BACON domains in bacteriophages by analysing over 2 million viral contigs. We found a high biodiversity of BACON in seven gut phage lineages, including five known crAss-like phage lineages and two novel gut phage lineages that are distantly related to crAss-like phages. In three BACON-containing phage lineages, we found that BACON domain tandem repeats were associated with phage tail proteins, suggestive of a possible role of these repeats in host binding. In contrast, individual BACON domains that did not occur in tandem were not found in the proximity of tail proteins. In two lineages, tail-associated BACON domain tandem repeats evolved largely through horizontal transfer of separate domains. In the third lineage that includes the prototypical crAssphage, the tandem repeats arose from several sequential domain duplications, resulting in a characteristic tandem array that is distinct from bacterial BACON domains. We conclude that phage tail-associated BACON domain tandem repeats have evolved in at least two independent cases in gut bacteriophages, including in the widespread gut phage crAssphage.

Introduction

Bacteriophage populations are essential for stability and proper functioning of the human gut microbiome[95]. Gut phages may provide immunity against bacterial pathogens[11], and changes in gut virome composition have been observed in diseases such as inflammatory bowel disease[66], type I diabetes[96], malnutrition[69] and colorectal cancer[97]. Although the role of phage populations in the gut microbiome is increasingly better understood, the role of individual phage species remains almost completely unknown. This last fact is exemplified by crAssphage, a phage that was first described from metagenomic datasets in 2014. Since then, crAssphage has been recognised as the prototypical member of an expansive family of phages[86] that is abundant in the human gut[65]. The crAss-like phages are distributed globally[62], present in approximately half the human population[62,83], and very abundant in some individual gut viromes[85]. However, little

is known about the role, evolution and distinguishing characteristics of specific crAss-like phage lineages.

One characteristic crAssphage protein (gp64 in NC_024711.1) encodes a tandem repeat of eight *Bacteroides*-associated carbohydrate-binding often N-terminal (BACON) domains[83]. BACON domains are most commonly found in carbohydrate-binding proteins of *Bacteroidetes* bacteria, but despite their name, a direct role in carbohydrate-binding for BACON domains is yet to be proven. The only currently published BACON domain X-ray crystallography structure in human gut *Bacteroidetes* indicated that they form links between carbohydrate-binding domains and are anchored onto the surface of *Bacteroides* cells and do not perform a carbohydrate-binding function themselves[98]. This X-ray crystallography structure also indicated that BACON domains form fibrous immunoglobulin-like (ig-like) β-sandwiches[98]. This matches the association between BACON domains and carbohydrate-binding proteins, as bacterial and viral ig-like domains are often associated with proteins possessing binding functions[99]. Phage ig-like domains are implicated in binding to phage hosts[100] and carbohydrates[101], and are common in surface proteins such as tail fibres. In the tail fibres and spikes of taxonomically diverse phages like T4[102,103], p2[104], HK620[105], p22[106,107] and Sf6[108], ig-like and other β-hairpin folded domains commonly form tandem repeats[109]. Likewise, in crAssphage the ig-like BACON domains form a tandem repeat.

Proteins with domain tandem repeats are found throughout the tree of life[110]. Because expansions of domain tandem repeats are a marker for adaptation to new environments, these structures are considered to be evolutionary hotspots[111]. Repeat expansions often involve duplications of multiple domains[112] that occur internally in the tandem array and not at the protein termini, as is common in non-tandem repeated domains[112]. Following tandem domain duplications, the independent units in a repeat rapidly evolve until only residues necessary for correct folding are conserved[113]. Once established in this manner, and contrasting their status as evolutionary hotspots, domain tandem repeats are often stable over long evolutionarily timespans[114]. As a result of their evolutionary mechanism, domain tandem repeats can be used to study the evolution and adaptational history of species. This is particularly true in eukaryotes, where they are well studied[110]. In the context of rapid and reticulate phage evolution, much less is known about domain tandem repeat evolution.

Here we studied the evolution of BACON domain tandem repeats in crAss-like phages. First, we constructed a profile hidden Markov model (HMM) of the crAssphage-like BACON domain (crAss-BACON), which we identified in half of the known candidate crAss-like phage lineages and in two novel gut phage lineages. Tandem repeats of the crAss-BACON were

exclusive to two crAss-like phage candidate genera and one of the novel lineages and were always genomically flanked by tail proteins. In phage lineages with single (non-repeated) crAss-BACONs, they were associated with the replicative and head regions of the phage genome. By studying crAss-BACON evolution, we showed that crAss-BACON tandem repeats are divergent from bacterial domains, whereas individually occurring crAss-BACON domains are more closely related to bacterial sequences, indicating frequent horizontal transfer between viral and cellular organisms. Finally, we showed that the crAss-BACON tandem repeat in the most widespread crAss-like phage lineage has evolved through multiple single-domain duplications.

Results and Discussion

Construction of a Specific Profile HMM of the crAss-like Phage BACON Domain

We started our study of BACON domain diversity and evolution by constructing a profile HMM distinctive to crAssphage-like BACON domains (crAss-BACONs). The *Bacteroides*-derived BACON domain that was available in Pfam (PF13004)[115] contained 4 highly conserved residues[112,113] out of a total of 45 (Figure 1a)[116]. These are an N-terminal tryptophan (position 11 in Figure 1a), a central asparagine and arginine (positions 27 and 33) and a C-terminal glutamine (position 52). To improve identification of crAss-BACONs with few conserved residues, we performed iterated distant homology searches (see Materials and Methods). This iterative search identified 605 BACON domains in 122 crAss-like contigs after four iterations. With these 605 domains, a crAss-BACON profile HMM was made. Comparison between the crAss-BACON profile HMM and PF13004 showed conservation of all four characteristic BACON residues (Figure 1a,b). The main difference between PF13004 and crAss-BACON profile HMM was a shorter N-terminus of the domain. The first 10 residues of the domain were absent altogether, while low occupancy scores in the profile HMM show that some of the 10 subsequent residues were absent from up to 30% of crAss-BACONs. High insert probabilities within the first 10 residues of both profile HMMs further show that this region of the domain is flexible.

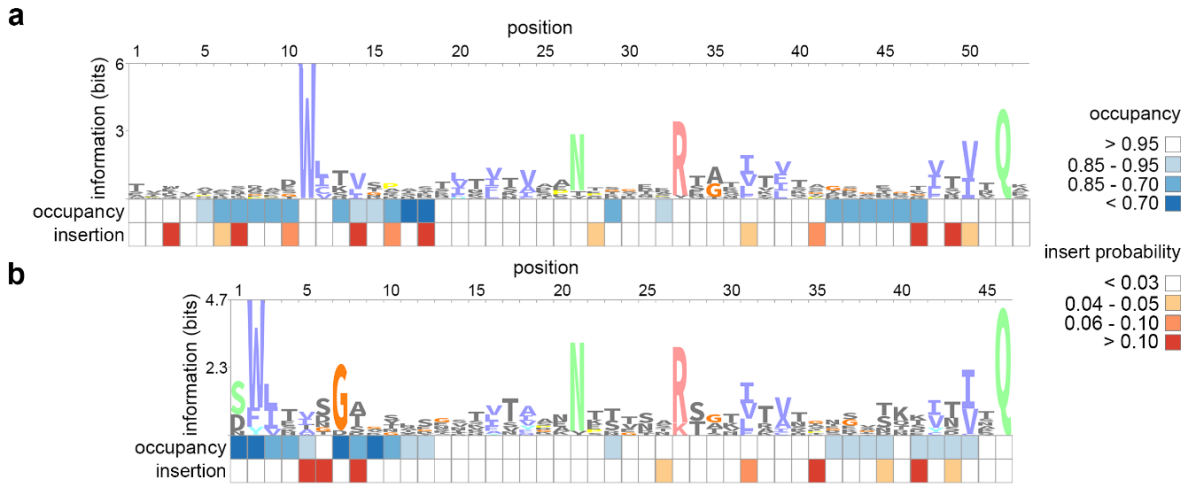


Figure 1. Sequence logos of Bacteroides-associated carbohydrate-binding often N-terminal (BACON) domain profile hidden Markov models (HMMs) show the divergence between (a) the bacterial BACON domain and (b) the crAssphage-like BACON domain (crAss-BACON). *Profile HMMs were constructed from (**a**) 304 Bacteroidetes domains (PF13004) and (**b**) domains identified in crAss-like phages. These sequence logos are representations of profile HMMs, which contain probability scores for each amino acid residue at each position in an alignment. In addition, profile HMMs contain probability statistics for insertions and deletions at each position. Occupancy scores denote the probability that an amino acid residue is found at a given position (i.e., low values mean a deletion is more likely to occur at that position). Insertion scores denote the probability of an insertion after a given location. Images were constructed with the Skylign webserver*[117,118].

The dataset in which we searched for BACON domains included contigs from an earlier study that presented a preliminary crAss-like phage taxonomy[85]. That study classified human gut-associated crAss-like phages into ten candidate genera and four candidate subfamilies. We identified crAss-BACONs in all four candidate subfamilies including *Alpha-*, *Beta-*, *Gamma-* and *Deltacrassvirinae*, but not in all ten candidate genera. Candidate genera I, II, III, V, VI and VII contained crAss-BACONs, whereas IV, VIII, IX and X did not. *Alphacrassvirinae* is the most widespread subfamily that contains candidate genus I and the prototypical crAssphage. This subfamily contains two candidate genera with crAss-BACONs (I and III) and two without it (IV and IX). BACON domains are particularly prevalent in candidate genus I, where 55 of the 63 genomes that were previously categorised in this genus contain crAss-BACONs. In candidate genus III, 7 of the 22 genomes contain crAss-BACONs. Likewise, *Deltacrassvirinae* contains one candidate genus with crAss-BACONs (VII, crAss-BACONs found in 10 of the 37 genomes) and two without it (VIII and X). We found crAss-BACONs in *Betacrassvirinae*

(candidate genus VI, crAss-BACONs in 3 of the 22 genomes), which contains the first isolated crAss-like phage[87], and in *Gammacrassvirinae* (II and V, crAss-BACONs in 13 of 14, and 15 of 18 genomes, respectively). Thus, crAss-BACONs were absent from some lineages, even though the search space contained genome sequences from all established candidate crAss-like phage genera that a previous publication showed were (near-)complete[85]. We conclude that crAss-BACONs are widespread but not universally conserved among the taxonomically diverse crAss-like phage lineages.

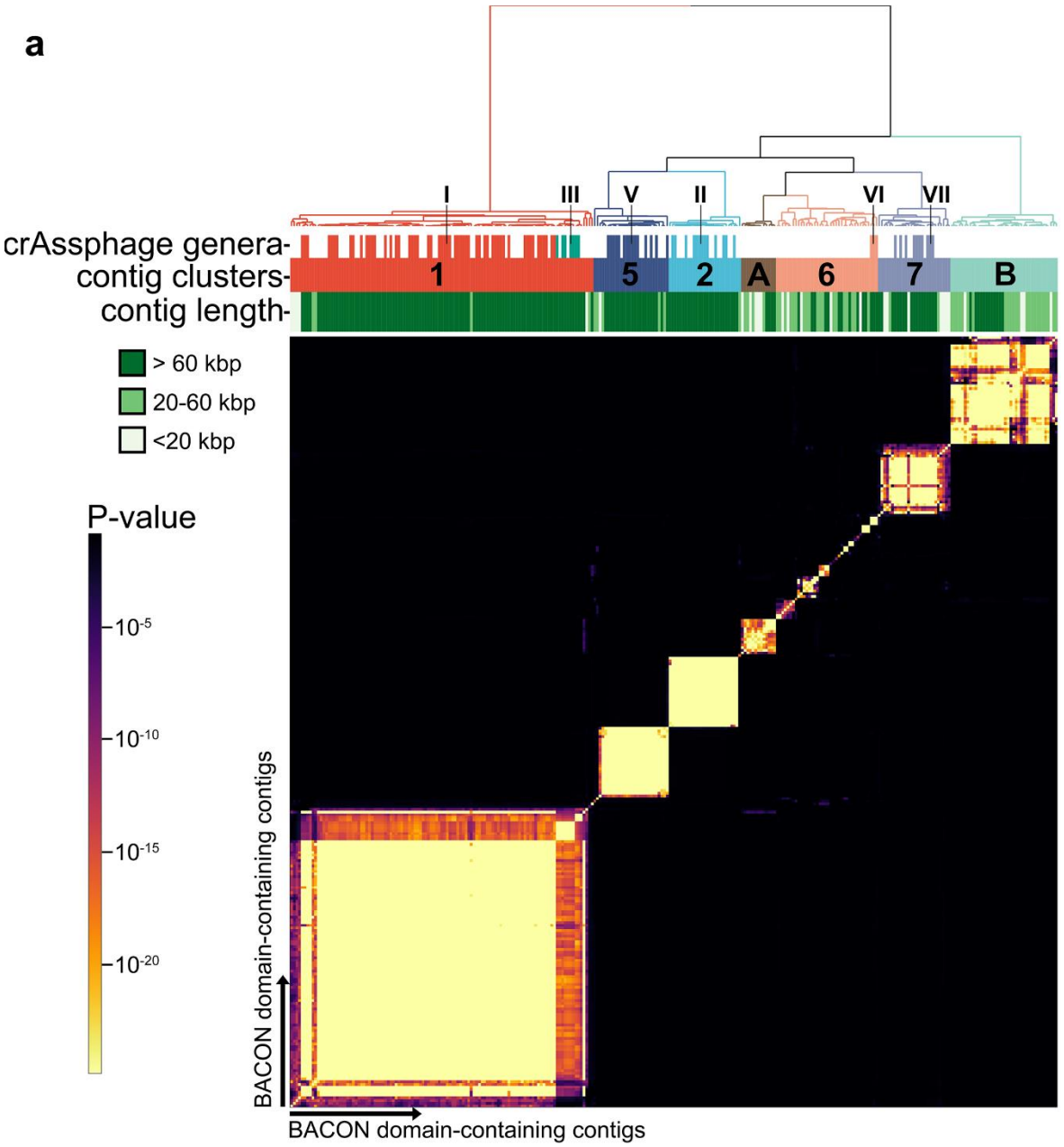
BACON Domains are Found in Diverse Phages

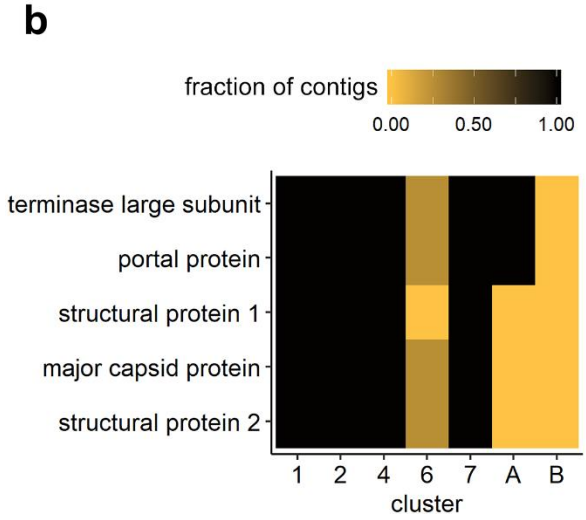
As BACON domains were found in taxonomically diverse crAss-like phages, we next determined how widespread they were in other phages. For this, we queried 2,147,193 viral sequences from seven metaviromics studies for the crAss-BACON profile HMM (Figure 1b) and identified 801 BACON domains in 246 contigs. Despite using a search space with viral contigs from a wide variety of ecosystems, contigs where crAss-BACONs were identified largely originated from human intestinal or sewage datasets. Using a profile HMM for sequence similarity searches allows for detecting considerably divergent sequences. Still, the crAss-BACON profile HMM was initially constructed from crAss-like phage sequences, so it is possible that very divergent BACON domains may have been missed. Additional biomes where crAss-BACON containing contigs were found include cow rumen (n = 2), chicken ceca (n = 2), sheep rumen (n = 1) and wombat intestines (n = 1), as well as three contigs originating from human oral metagenomes[119] (see Table S1, Supplementary Materials).

To investigate the biodiversity of the organisms containing crAss-BACONs, we quantified the similarity of these contigs through their shared protein family content[120–122]. Hierarchical clustering of significantly shared protein family content per contig pair resulted in seven distinct clusters (Figure 2a). Five clusters contained contigs that were previously proposed as candidate crAss-like phage genera[85]. Cluster 1 represents the *Alphacrassvirinae* candidate subfamily and contains contigs from candidate genera I and III. Clusters 2, 5 and 7 conform with candidate genera II, V and VII, respectively. As previously observed for candidate genus VI[85], cluster 6 is highly diverse (see also Figure S1, Supplementary Materials). The remaining two clusters, which we labelled A and B, contain contigs that were not previously identified as crAss-like phages.

We next determined how the seven clusters were related by performing a phylogenetic analysis of five crAss-like phage head proteins that are conserved in all crAss-like phages[86]. To maximize correct phylogenetic inference, we selected only contigs over 85 kbp for this analysis. The five crAss-like phage head proteins included a terminase and a major capsid protein, often among the most conserved phage proteins[123,124]. Conversely to this and the conservation of these proteins in known crAss-like phages, one of the head proteins was absent from all cluster 6 contigs, whereas the other four were present in a quarter of the contigs in this cluster (Figure 2b). Their absence likely suggests that cluster 6 includes multiple phage lineages and/or incomplete genome fragments. The five head proteins were also partially or completely absent from clusters A and B. Cluster A contigs contained only the terminase and portal proteins, while cluster B contained no homologs to any of the five head proteins (Figure 2b). This suggests that cluster A is more closely related to crAss-like phages than cluster B. A phylogenetic analysis of the concatenated multiple sequence alignment of the two head proteins, which had homologs in all clusters except B, revealed that clusters 7 and A were the most distant lineages (Figure 2c). Considering that all five head proteins are present in cluster 7 but not in cluster A, we suggest that cluster 7 is more closely related to the other crAss-like phages than cluster A.

To further analyse the phage lineages in clusters A and B, we functionally annotated the protein-encoding genes on the longest sequences from clusters A and B. These contigs were 154,123 bp (cluster A) and 87,116 bp (cluster B) long and contained 180 and 110 ORFs respectively (see Table S3 and S4, Supplementary Materials for annotation tables). Of the 180 ORFs in the cluster A representative contig, 96 had significant hits to the nr database. Most of these hits (64 hits) were found in *Bacteroidetes* species. Ten of the first fifteen genes on this contig hit genes in a *Parabacteroides merdae* genomic region. As the seventeenth gene on the *P. merdae* contig is a transposase, this region likely signifies a partial prophage[125]. This assertion is strengthened by the presence of two phage antirepressor proteins, which act in prophage induction[126]. About half the ORFs (56 out of 110) in the representative cluster B contig hit *Proteobacteria* proteins, specifically from *Acinetobacter baumannii* (35 ORFs) and *Klebsiella pneumoniae* (14 ORFs). Multiple cluster B ORFs hit phage structural proteins located in bacterial genomes, including three in *A. baumannii,* one in *K. pneumoniae*, one in *Desulfovibrio alaskensis* and one in *Rhodobacteriaceae.* Like the cluster A contig, the cluster B contig may thus have a temperate lifestyle. The high number of ORFs with hits in *Proteobacteria,* coupled with only three ORF hits to *Bacteroidetes* species, could indicate that this cluster B phage infects *Proteobacteria,* a different phylum than the crAss-like phages[86,87].


a
I
III
V
II
VI
VII
crAssphage genera-
contig clusters-
contig length-
1
5
2
A
6
7
B
> 60 kbp
20-60 kbp
<20 kbp
P-value
10-5
10-10
10-15
10-20
BACON domain-containing contigs
BACON domain-containing contigs


b
fraction of contigs
0.00
0.50
1.00
terminase large subunit
portal protein
structural protein 1
major capsid protein
structural protein 2
1
2
4
6
7
A
B
cluster

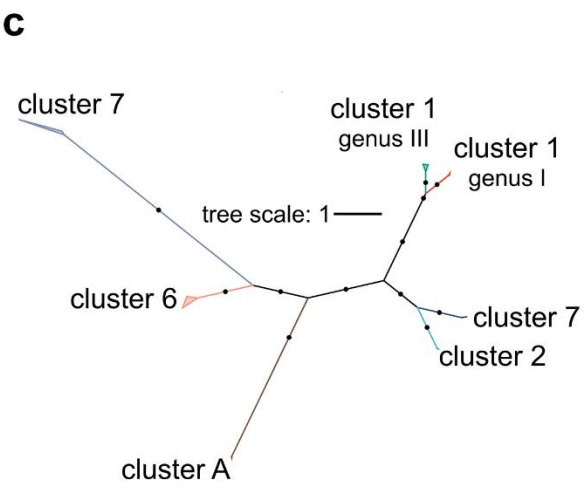

c
cluster 7
cluster 1
genus III
cluster 1
genus I
tree scale: 1
cluster 6
cluster 7
cluster 2
cluster A

Figure 2. crAss-BACON-containing contigs form seven distinct clusters, some of which are distantly related to crAss-like phages. (A) *Heatmap of crAss-BACON-containing phage contigs with seven clusters denoted. Clustering was performed based on a hypergeometric* p*-value calculated from the number of shared homologous ORFs between each pair of contigs (see Materials and Methods). Marks in the crAssphage genera row indicate contigs which have previously been identified as members of one of the ten candidate crAss-like genera. Roman numerals indicate the candidate crAss-like genera to which contigs belong*[85]*.* ***(B)*** *Variable presence of conserved crAssphage-like head proteins among the seven clusters. The fraction denotes the fraction of contigs in the cluster that contained that protein. No homologs were found from 1/5 proteins in cluster 6 contigs, 3/5 in cluster A and 5/5 in cluster B.* ***(C)*** *Approximate maximum likelihood tree of crAss-like terminase and portal protein homologs. Black dots indicate branches with ultrafast bootstrap support >90. Names of crAss-like phage candidate genera are according to Guerin et al*[85]*. No clade with cluster B contigs is present due to the absence of crAss-like terminase and portal proteins from the contigs of this cluster.*

If this were to be confirmed, it would mean that BACON domains are not as exclusive to the *Bacteroidetes* as previously thought[116]. In contrast, the WIsH host-prediction algorithm (published accuracy of 60%) did predict *Bacteroidetes* hosts for cluster B contigs, consistent with crAss-like phages[127]. These conflicting results highlight the importance of experimental research to make firm conclusions about phage host range[60,128].

Few ORFs in either the cluster A or B contigs are homologous to phage proteins in the RefSeq database (see Table S3 and S4, Supplementary Materials). Of the two, cluster A had the most hits to known phages. These included 12 ORFs with homology to *Cellulophaga* phages phi17:2 and phi4:1 proteins, which represented roughly 10% of the ORFs predicted on the cluster A contig. Since *Cellulophaga* phage phi14:2 is a distant relative of crAssphage[86], this is a further indication that cluster A is distantly related to crAss-like phages. Only five cluster B ORFs showed direct similarity to proteins from isolated phages (BLASTp *E*-value $<10^{-5}$). These phages infect four different phyla and are unrelated to crAss-like phages. The above results show that cluster A and B genomes likely represent newly described phage lineages.

BACON Domains Have Diverse Configurations in Phages

After establishing that crAss-BACONs are widespread in crAss-like phages and beyond, we next looked at differences in crAss-BACON architecture among the phage clusters. Domain

tandem repeats in cellular organisms tend to rapidly evolve following domain duplications, after which they remain stable and conserved within species[114]. The tandem repeat of eight crAss-BACONs per ORF, as found in the prototypical crAssphage[83], was unique to cluster 1 (Figure 3a), where the crAss-BACON sequences were highly conserved (Figure S2, Supplementary Materials).

Other crAss-BACON architectures were found outside of cluster 1 where all crAss-BACON-containing ORFs had fewer than eight domains. Most crAss-BACON ORFs in clusters 2, 5 and 6 contain a single domain (Figure 3a). Multiple contigs in clusters 2 and 5 possessed multiple crAss-BACON ORFs (Figure 3b), which in some cases were separated by several other ORFs (Figure 3c). Like cluster 1, most crAss-BACON ORFs in clusters 7, A and B contained more than one domain (Figure 3a). While crAss-BACONs in cluster A ORFs were not organized in a tandem array, clusters 7 and B contained instances of domain tandem repeats of five and six domains. Therefore, crAss-BACON tandem repeats are limited to clusters 1, 7 and B.

Because repeats of β-hairpin folds such as BACON are common in phage tails[109], we hypothesised that the crAss-BACON ORFs in cluster 1, 7 and B may encode tail-associated proteins. The poor conservation of the C-terminus of cluster 1 crAss-BACON ORFs (Figure S2) might be consistent with their role as tail fibres, as phage receptor-binding proteins commonly have low conservation in one terminus[129]. We could not directly verify the function of crAss-BACON tandem repeat-containing ORFs, as only two cluster 2 crAss-BACON ORFs had sequence similarity to proteins from the nr database with functional predictions. These two hits, to a putative glycosyl hydrolase (WP_116811532.1, *E*-value of 6×10^{-8}) and a putative T9SS sorting domain-containing protein (WP_091894873.1, *E*-value of 9×10^{-8}), were solely based on an alignment to BACON domains. To circumvent this lack of detectable sequence similarity of crAss-BACON ORFs to proteins with predicted function, we instead analysed their genomic neighbourhoods (Figure 3c). This showed that ORFs with crAss-BACON tandem repeats in clusters 1, 7 and B are located near predicted tail or tail fibre proteins (Figure 3c). The crAss-BACON ORFs from cluster 1 in particular are flanked by multiple predicted genes encoding phage tail proteins. In clusters 2, 5, 6 and A, the crAss-BACON-containing ORFs are in proximity of proteins with a variety of other predicted functions. In cluster 2, we identified multiple crAss-BACON ORFs that were located on either side of the head section of the genome. Upstream of the head section were crAss-BACON ORFs and T9SS sorting domain-containing proteins. In one case, a crAss-BACON ORF also contained a T9SS sorting domain. The T9SS secretion system, which is only found in *Bacteroidetes*, is used for transport of

virulence factors, nutrient acquisition and gliding motility[130]. How this relates to the presence of BACON domains and their correlation to carbohydrate-binding requires additional research.

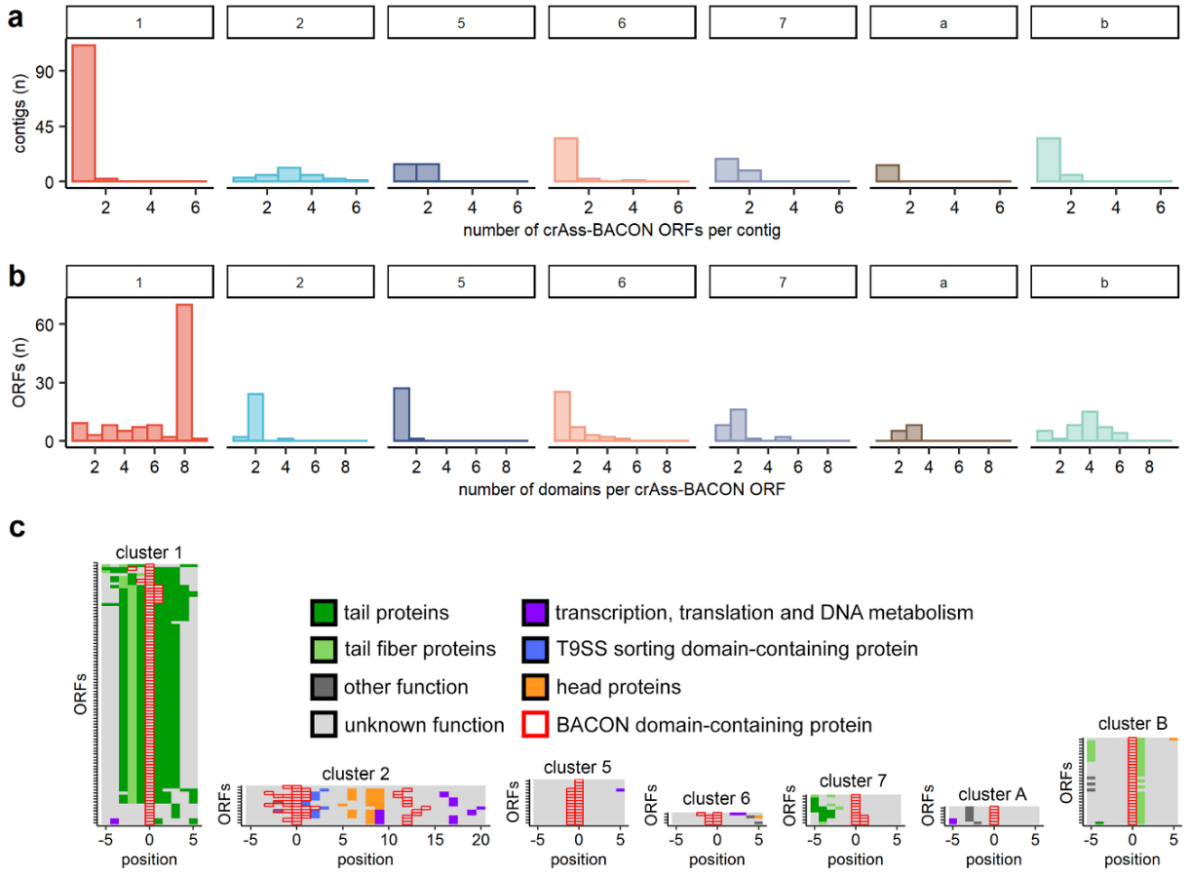


Figure 3. BACON-containing ORFs have diverse domain architectures and genomic neighbourhoods. *Histograms of* ***(A)*** *the number of BACON domains per ORF and* ***(B)*** *the number of BACON domains containing ORFs per contig show the diversity in domain architecture between clusters.* ***(C)*** *The genomic neighbourhood of crAss-BACON ORFs. Note that the ORFs containing crAss-BACON tandem arrays (clusters 1,7 and B) are flanked by tail proteins. Only contigs with more than 10 ORF predictions and at least one homolog with a predicted function other than "hypothetical protein" are shown (cluster 1:* n *= 88/113 contigs are shown, 2:* n *= 13/27, 5 = 15/28, 6 = 4/38, 7 = 10/27, A = 6/13, B = 29/40).*

Annotation of the largest contigs in all clusters (Figure 4) indicated that in clusters 2, 5 and 6 the crAss-BACON ORFs are located near head and structural proteins, whereas cluster A crAss-BACON ORFs are located in the transcriptional section of the genome. Considering that crAss-BACON ORFs in clusters 2, 5 and 6 are located among head and structural proteins, they may be phage capsid-decorating binding proteins[11,131]. As only cluster 1, 7 and B ORFs with crAss-BACON tandem repeats are located in the tail section of the genome, these seem to represent

examples of the β-hairpin folded repeats that are common in phage tail fibres and spikes[109]. Structural analysis revealed that bacterial BACON domains are linkers that anchor carbohydrate-binding domains to the *Bacteroides* cell surface[98]. Thus, we hypothesised that the tail-associated crAss-BACON tandem repeats may have been recruited by crAss-like phages to bind to the *Bacteroides* cell surface, and that the evolutionary expansion of the tandem repeats that are specific to tail-associated crAss-BACON ORFs may have enhanced the binding capability of the phage tails.

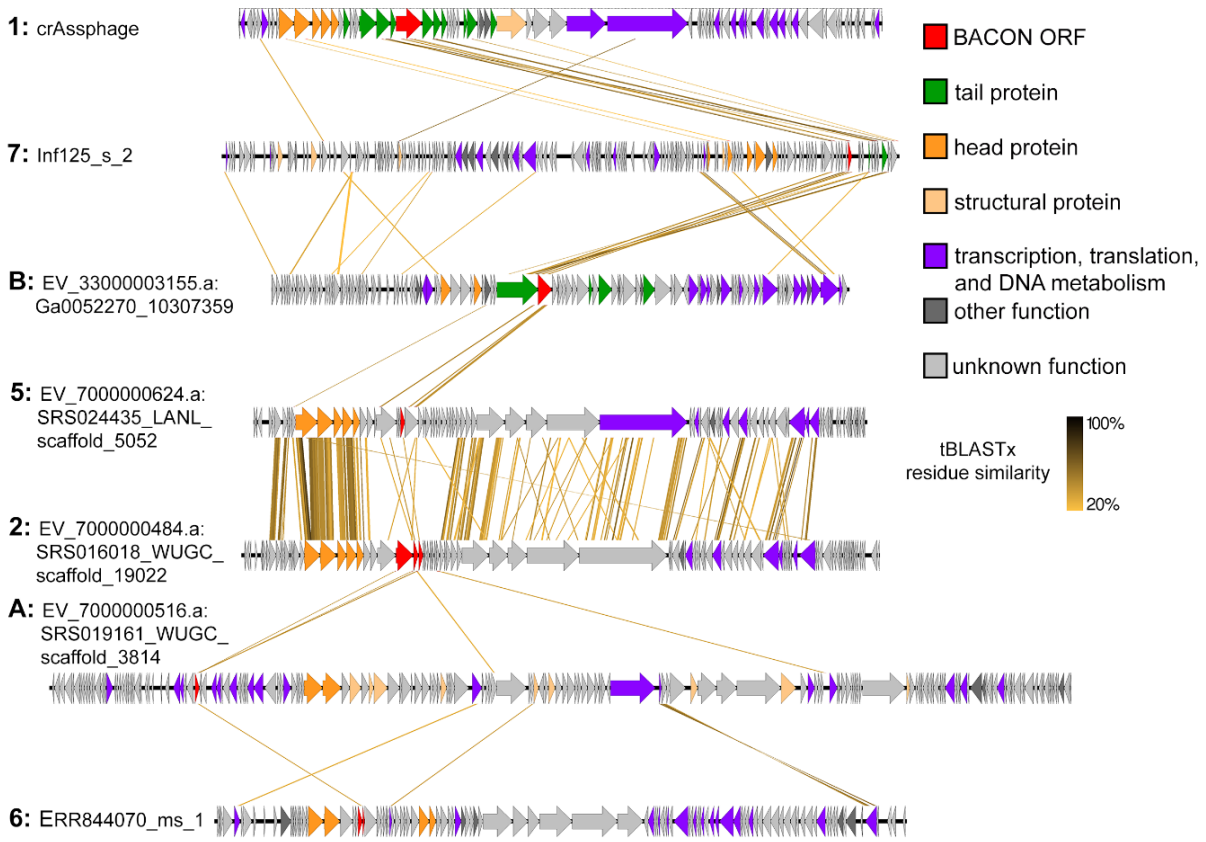


Figure 4. Whole genome comparisons of selected contigs from each phage cluster show that horizontal transfer has occurred in the neighbourhood of crAss-BACON ORFs between the otherwise unsimilar clusters 1, 7 and B. *crAss-BACON tandem repeated proteins in these clusters are located in the tail section of the genome. In all clusters except from B, head proteins are those proteins that were identified in the phylogenetic analysis in Figure 2c.*

Recurrent Evolution of BACON Domain Tandem Repeats in Phages

As shown above, crAss-BACON tandem repeats are associated with tail proteins in taxonomically distinct phages. Next, we focused on how these repeats evolved. In phage clusters where crAss-BACON tandem repeats were found (1, 7 and B), the genomic regions

that contained crAss-BACON ORFs had high sequence similarity, whereas the rest of the phage genomes did not (Figure 4). This suggests that horizontal transfer of the region containing the crAss-BACON ORF has occurred between these phage lineages and their hosts within the crAss-BACON ORFs, as has been previously observed for phage immunoglobulin domains[100]. Vertical transfer of the crAss-BACON ORFs cannot fully explain the similarity of crAss-BACON domains between these phages. The genomes of clusters 1, 7 and B show very little similarity, as shown in Figure 4 and as exemplified by the low conservation or even absence of genes such as the terminase (Figure 2b).

To investigate BACON domain evolution in phages and bacteria, we created a phylogeny of 1509 BACON domains (Figure 5a). This included all 1205 crAss-BACON domains found here and the 304 bacterial BACON domains from which PF13004 was made. Figure 5b–d depicts the crAss-BACON phylogenetic tree with colours indicating clades that contain domains from the crAss-BACON tandem repeats found in clusters 1, 5 and B, respectively. In cluster 1, domains that occupy a specific position in the crAss-BACON tandem repeat (e.g., first domain from the N-terminus, referred to with numbers in Figure 5b–d) each formed a distinct clade (Figure 5b). As domains that occupy a specific position in the cluster 1 crAss-BACON tandem repeat form distinct clades, with no overlap between them, the domain order of the arrays is conserved. This illustrates the evolutionary stability of the cluster 1 crAss-BACON tandem repeat. The clade that contains the first domain in the cluster 1 array (counted from the N-terminus) also contained a number of bacterial domains, whereas no bacterial domains are found in the clades of the seven other domains (Figure 5a). This reveals that this crAss-BACON tandem array resulted from domain duplications within the cluster 1 phage lineage. As the first domain is more closely related to bacterial domains than the other domains in the tandem array, it may be the ancestral domain from which the array expanded. The tree contains two branches that contain cluster 1 crAss-BACON domains that occupy positions 1, 7 and 8 in the array, and those at positions 3, 4, 5 and 6. This means that, similar to domain tandem repeat expansions in cellular organisms[112], cluster 1 crAss-BACON expansion occurred by internal duplications, rather than exclusively at the N- and C-termini (Figure 5b). However, unlike cellular organisms[112], cluster 1 crAss-BACON expansion involved several single domain duplications instead of simultaneous duplication of multiple domains (Figure S3, Supplementary Materials).

The crAss-BACONs in the cluster 7 and B tandem arrays showed a greater spread in the tree (Figure 5c,d). Like cluster 1, these two clusters each contained one crAss-BACON that is closely related to bacterial domains (domain 4 in cluster 7 and domain 6 in cluster B). A further

domain from each was closely related to cluster 1 crAss-BACONs (2/7 and 4/B), which seems to be the result of horizontal domain transfer. Each cluster also contained one domain in an isolated branch (5/7 and 3/B) and a number of domains that formed a separate branch (1/7, 3/7, 1/B, 2/B, 5/B). These latter domains resulted from duplications, whereas those that were closely related to bacterial domains may be the ancestral domains. The internal divergence of the domains in these tandem repeats as well as the frequent horizontal transfer of BACON domains between phages and bacteria make it difficult to fully ascertain their evolutionary history.

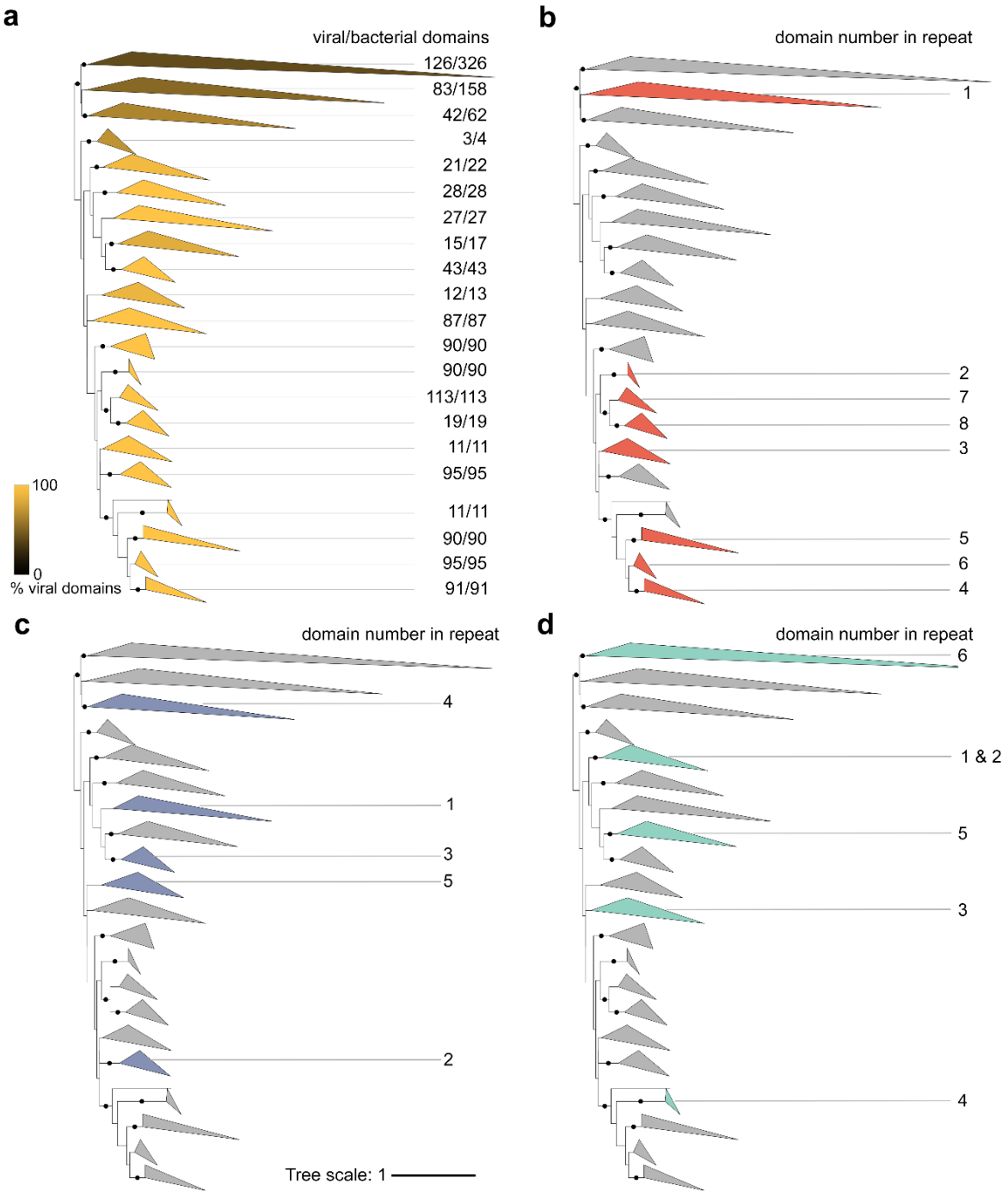


Figure 5. Domain phylogeny of crAss-BACONs in bacteria and phages reveals that the cluster 1 crAss-BACON tandem repeat resulted from multiple duplication events, whereas those in clusters 7 and B resulted from both horizontal transfer and duplications. *Displayed is an unrooted approximate maximum likelihood tree[132] of 1205 crAss-BACONs and 304 BACON domains from Bacteroidetes species (those that were used to produce PF13004). Dots on branches denote bootstrap support >90. The four versions of the tree depict the fraction of crAss-BACONs out of the total number of domains in each clade* ***(A)****, and the locations of domains from the crAss-BACON tandem repeats in cluster 1* ***(B)****, 7* ***(C)*** *and B* ***(D)****. Shading and numbers in* ***(A)*** *denote the fraction of viral domains in each clade. Numbered clades in* ***(B)****,* ***(C)*** *and* ***(D)*** *represent the positions of domains in the crAss-BACON tandem repeat as counted from the N-terminus. For uncollapsed tree, see Figure S4, Supplementary Materials.*

Our domain phylogeny uncovered an extensive evolutionary history of domain duplications that is characteristic to cluster 1 crAss-BACON ORFs. Identification of additional sequences from clusters 7 and B may further clarify the evolution of crAss-BACON tandem repeats in these clusters. As the cluster 1 crAss-BACON ORF is likely a tail protein, elucidation of its function may provide insight into the host interactions of the most widespread gut phage. The first crAss-like phage that was recently isolated (φcrAss001[87]) belongs to the heterogeneous candidate genus VI and not to the *Alphacrassvirinae* candidate subfamily that contains the longest and most conserved crAss-BACON tandem array (cluster 1, see Figure 2 and Figure S1). Interestingly, our distant homology searches did not identify any crAss-BACONs in the φcrAss001 genome. Determination of the function of the tail-associated crAss-BACON ORF thus awaits the isolation of crAss-like phages from clusters 1, 7 or B.

Conclusions

We investigated the evolution of BACON domain tandem repeats in crAssphage and other gut bacteriophages. We demonstrated that crAss-like phage BACON domains (crAss-BACONs) are widespread in the crAss-like phage family, and we identified crAss-BACONs in two novel gut phage lineages. We further found that crAss-BACON tandem repeats are associated with tail fibres in three gut phage lineages. In two lineages (clusters 7 and B), these repeats are the result of horizontal transfer and tandem duplication. The third lineage with tail-associated crAss-BACON tandem repeats includes the prototypical crAssphage (cluster 1, candidate

subfamily *Alphacrassvirinae*). In this lineage, we showed how a stable tandem repeat of eight crAss-BACONs has resulted from multiple single domain duplication events. The presented results show that focus on uncharacterised proteins can provide insight into the enormous biodiversity and evolutionary dynamics of the viral world.

Materials and Methods

Data

We identified BACON domains across a broad range of viruses using two datasets that included both genomic and metagenomic sequences. The first dataset consisted of 661 crAss-like phage sequences from five recent publications[62,83–86] that ranged in size from short fragments of 623 bp to complete and near-complete genomes of up to 104,752 bp (median: 27,648 bp). The second dataset was more diverse and consisted of 2,147,193 viral contigs from four recent viromics papers[36,119,133,134]. The contigs in the second dataset originated from ecosystems across the biosphere, including host-associated microbiomes from humans and other animals, fresh and saltwater aquatic ecosystems, and marine and terrestrial sediments (see Table S1, Supplementary Materials, for metadata on contigs in which crAss-BACONs were identified). A profile hidden Markov model (HMM) of the BACON domain (PF13004), constructed from 304 protein domains present in *Bacteroidetes* bacteria, was downloaded from the Pfam database v32.0 on 1 April 2019[115]. The 304 bacterial BACON domain sequences, from which PF13004 was constructed, were also retrieved from Pfam.

Identification of BACON Domains

All bioinformatics tools mentioned below were run with default settings and cut-offs, except where stated explicitly.

Before searching for BACON domains, we predicted all ORFs in both datasets of viral sequences using Prodigal v2.6.1[135] with translation table 11. BACON domains were identified in the dataset of crAss-like phage sequences using iterative searches with hmmsearch v3.1b2[136]. In the first iteration, PF13004 was used as profile HMM for the search against the crAss-like phage dataset. For subsequent iterations, profile HMMs were constructed using hmmbuild v3.1b2 with the aligned domain hits of the previous iteration as input. A total of four iterations were performed, at which point the number of hits converged. The final dataset of crAss-BACONs contained 605 domains in 171 ORFs on 122 contigs. A crAss-BACON profile HMM

was made with these domains as above, which was subsequently used to search in the metagenome dataset with hmmsearch v3.1b2[136]. This search identified an additional 801 BACON domains in 296 ORFs on 246 contigs.

Clustering of BACON Domain-Containing Viral Contigs

To obtain a tentative taxonomical classification of the crAss-BACON-containing contigs, we clustered them based on shared protein content[120–122]. To increase clustering reliability, only contigs with more than 10 ORFs were retained. This decreased the dataset to 1205 BACON domains in 376 ORFs on 286 contigs. This dataset was used for all subsequent analyses. For a full overview of all domains, ORFs and contigs in this dataset and related metadata, see Table S1 and S2, Supplementary Materials. Homologous gene clusters were made by performing a BLASTp all versus all on every ORF from every crAss-BACON-containing contig using Diamond v0.9.25[137]. The output was filtered for hits with a bit score >50. Subsequently, homologous ORF clusters were formed using the Markov cluster (MCL) algorithm with inflation parameter 2 (option -I 2)[138]. The number of shared homologous gene clusters between each contig pair was determined, from which the significantly shared gene content was calculated using the R function phyper. A Euclidean distance matrix of the hypergeometric *p*-values was constructed with the R function dist and used for clustering with the Ward.D2 algorithm as incorporated in R. The optimal number of clusters was determined using the NbClust v3.0 R package[139], whereas silhouette plots for the clusters were obtained using the factoextra v1.0.5 R package. Heatmaps were plotted using the heatmaply v0.16.0 R package[140].

To analyse the conservation of the cluster 1 crAss-BACON tandem repeat, all cluster 1 crAss-BACON ORFs with eight domains were selected and aligned with Clustal Omega v1.2.1[141]. Information content and residue conservation scores per position of the alignment were calculated using Geneious v.9.1.8[142].

Analysis of Novel Contig Clusters

Most crAss-BACON-containing contig clusters contained members of previously defined candidate crAss-like phage subfamilies and genera, except clusters A and B which consisted of highly dissimilar viral sequences (see Results and Discussion). To provide an initial sequence-based characterization of these novel viruses, the largest contigs from each cluster were extracted and annotated using the PROKKA v1.11 software tool[143], with the metagenomics option enabled in two separate runs for both the viral and bacterial settings (i.e., with --

metagenome and both --kingdom Bacteria and --kingdom Viruses options). Whole genome comparisons between the largest contig of each cluster were made using Easyfig v2.2.3[144], which performs tBLASTx searches along the entirety of two sequences. Easyfig employed the tBLASTx function of BLAST+ v2.9.0[145].

To further study the relation of the newly described contig clusters to known crAss-like phages, we performed a phylogenetic analysis of crAss-like phage head proteins. Five crAss-like phage head proteins that were previously used to determine crAss-like phage taxonomy[86] were extracted from the crAssphage genome (locus tags gp73–77 in NC_024711.1). They were used as queries for jackhmmer searches[136] against a database consisting of all predicted ORFs from the 166 crAss-BACON-containing contigs longer than 85 kbp. Analysis of the jackhmmer output showed that two of the five head proteins (gp76 and 77) had homologs in all but one of the phage clusters. These two proteins were selected for phylogenetic analysis. A separate alignment was made for each head protein using Clustal Omega v1.2.1[141]. The alignments of the head proteins were concatenated and positions with more than 95% gaps were removed using trimal v1.2[146] (option -gt 0.05). As was recommended in a recent meta-analysis of phylogenetic tree reconstruction tools[147], ten maximum likelihood trees were made using IQ-Tree v1.6[132] using model finder[148] and 1000 iterations of both the SH-like approximate likelihood ratio test and the ultrafast bootstrap approximation (UFBoot)[149] (i.e., options -alrt 1000 -bb 1000). Out of the ten constructed trees, the one with the highest likelihood (for which model finder selected VT+F+R8) was visualised using interactive Tree of Life v4.4.2[150].

Hosts were predicted for all crAss-BACON contigs with a length over 75,000 bp using the host prediction algorithm WiSH v1.0[127]. This program uses k-mer profiles of bacteria and phages to calculate likelihood scores for a given phage–host interaction, with the highest log-likelihood score denoting the most likely host. As bacterial genomes, we used 2613 complete bacterial genomes that were extracted from the PATRIC database that was downloaded on 20 June 2019[151]. A single genome sequence was selected for every genus in the database, with the highest score according to the formula C – 5 * M, where C and M are the CheckM completeness and contamination scores, respectively[152]. In case of a draw, the genome with the highest coarse consistency score was selected. These values were provided by the PATRIC database.

Examining of BACON ORF Genetic Neighbourhoods

To study the genomic neighbourhood of crAss-BACON ORFs, these plus five ORFs up- and downstream were collected. All ORFs were queried against the NCBI non-redundant protein

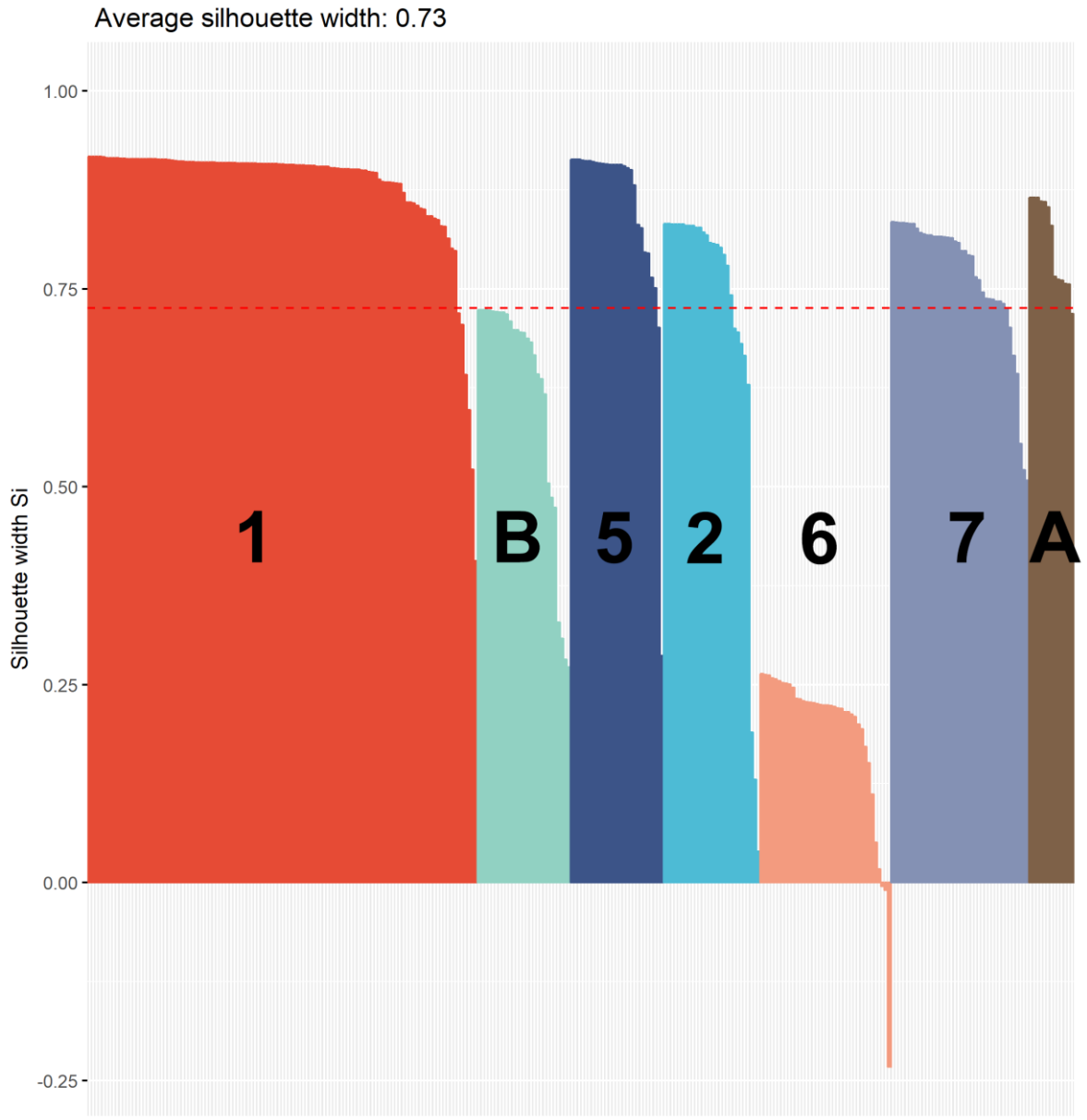
sequences collection[153] with BLASTp on the BLAST webserver[154] on 26 June 2019. In some contigs, the neighbourhood search only returned significant similarity (*E*-value $\leq 10^{-5}$) to proteins with a "hypothetical protein" function description. Because this provided no additional insight into the role of these 121 crAss-BACON ORFs, they were discarded from the analysis. All other neighbourhoods were plotted using the ggplot2 v3.1.0 R package.

Phylogenetic Analysis of BACON Domains

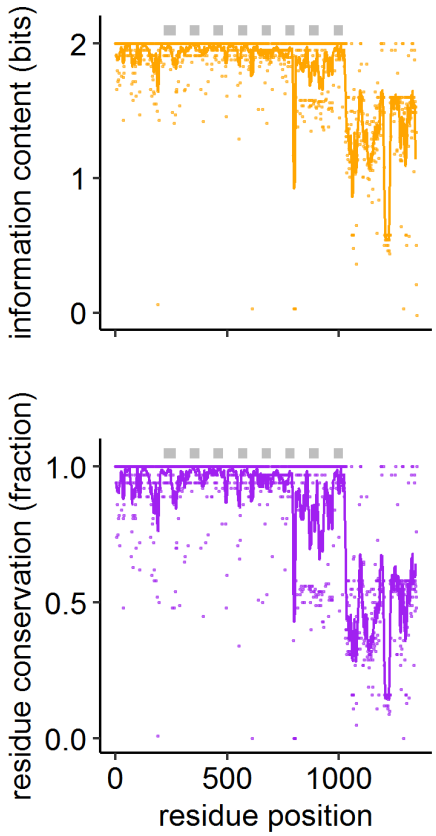
To study crAss-BACON evolution, we constructed an approximate maximum likelihood tree. CrAss-BACONs were first aligned with Clustal Omega v1.2.1[141]. While crAss-BACONs are homologous, they possess relatively low sequence similarity, as is characteristic for tandem domain repeats[113]. To improve likelihood and bootstrapping support of the phylogenetic analysis, we trimmed positions with more than 60% gaps using trimal v1.2[146] (option -gt 0.4). The resulting alignment maintained all four highly conserved crAss-BACON residues (see Results and Discussion). The tree was subsequently constructed as described above in Section 2.4. Out of the ten constructed trees, the one with the highest likelihood (for which model finder selected WAG+R5) was visualised using interactive Tree of Life v4.4.2[150].

To analyse whether the cluster 1 crAss-BACON array had evolved through simultaneous duplication of multiple domains, the eight domains from crAssphage were subjected to a BLASTp all versus all using BLAST+ v2.9.0[145], and bitscores were plotted, as has been described previously[112].

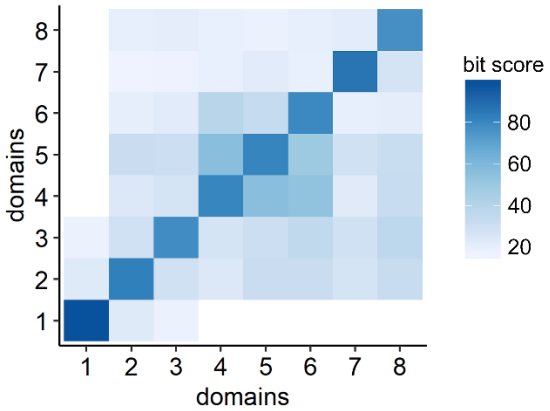
Supplementary Figures



Supplementary Figure 1: Silhouette plot[155] of the seven crAss-BACON-containing contig clusters shows that cluster 6 is weakly clustered due to heterogeneity (see also Figure 2a). *Each bar represents one contig, while silhouette width (y-axis) denotes the similarity of that contig to other contigs in the same cluster. Clustering quality decreases with decreasing silhouette width.*

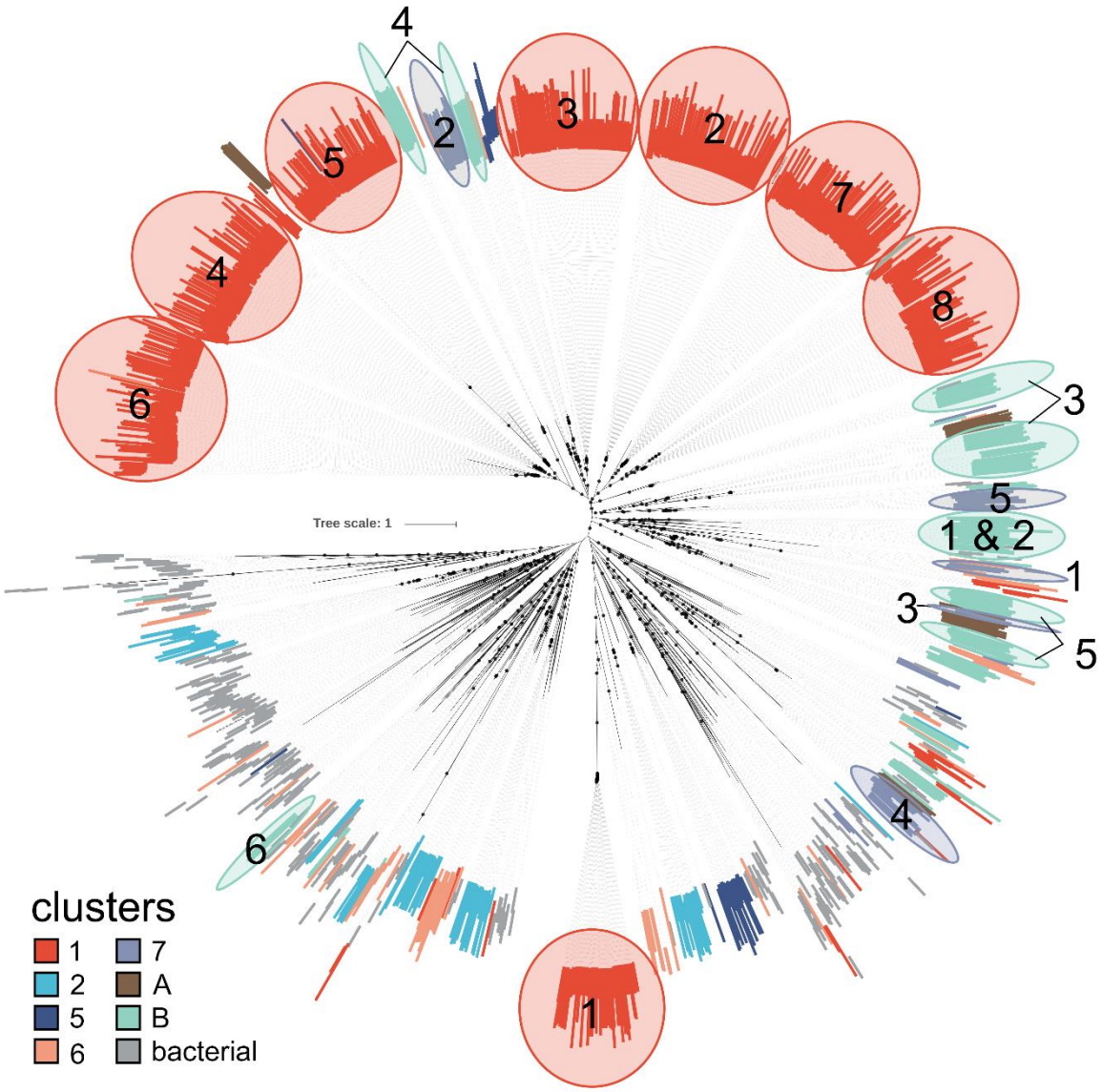


Supplementary Figure 2: Cluster 1 crAss-BACON ORFs are conserved, except for their C-termini. A Clustal Omega[141] multiple sequence alignment was made of all BACON ORFs from cluster 1 that contained eight domains. *The top graph represents the likelihood that the position contains a residue, the bottom graph represents the fraction of sequences which has a similar residue. Points are the individual positions, lines are a moving average with a period of 10. Grey blocks at the tops of the graphs are the positions of crAss-BACONs.*



Supplementary Figure 3: A similarity matrix of crAssphage BACON tandem repeats shows that the tandem array expansion occurred through single domain duplication events. *Bit scores are the result of a BLASTp all versus all search with crAssphage BACONs. The domain*

on the x-axis is the query, the domain on the y-axis the hit. Duplication of multiple protein domains at once would result in a checkerboard pattern[112]. Since no such pattern is evident, we conclude that the duplication events involved single domains. Also evident is that the first domain is divergent from the other seven, and that domains 4-6 duplicated internally.



Supplementary Figure 4: The same approximate likelihood tree as depicted in Figure 5a, but without collapsed branches.

3

Molecular and evolutionary determinants of bacteriophage host range

"Yet such is oft the course of deeds that move the wheels of the world:
small hands do them because they must, while the eyes of the great are elsewhere."
The Fellowship of the Ring; J.R.R. Tolkien

Published as: de Jonge, P.A.; Nobrega, F.L..; Brouns, S.J.J.; Dutilh, B.E.; *Molecular and Evolutionary Determinants of Bacteriophage Host Range*; Trends in Microbiology; 2019; 27; 51-63.

Abstract

The host-range of a bacteriophage is the taxonomic diversity of hosts it can successfully infect. One of the central traits to understand phages, host-range is determined by a range of molecular interactions between phage and host throughout the infection cycle. While many well-studied model phages seem to exhibit a narrow host-range, recent ecological and metagenomics studies indicate that phages may have a range of specificities, ranging from narrow to broad host-range. There is a growing body of work studying the molecular mechanisms that allow broad host-range phages to infect multiple hosts. These mechanisms and their evolution are of considerable importance to understanding phage ecology and various clinical, industrial and biotechnological phage applications. Here we review knowledge of the molecular mechanisms that determine host-range, provide a framework defining broad host-range in an evolutionary context, and highlight areas for additional research.

The importance of host-range

One hundred years after their initial discovery, research into bacteriophages, viruses that infect bacteria, is undergoing a surge of interest. This increased research interest is driven by a new appreciation of the ecological impact[134], potential applications[156], and role in human health[65,66] of phages. For example, metagenomic surveys have uncovered a wealth of viral diversity in natural and human-associated samples[65,157,158]; novel applications of phages are being developed (Box 1); and novel theoretical models have increased the focus on the eco-evolutionary roles of phages in natural microbial systems[159,160] (Box 2). To place the plethora of new information into context, it is crucial to understand the interactions that phages have with their hosts. Host-interaction has implications for how phages control microbial populations[161,162], how they facilitate horizontal gene transfer[49], and how they can be deployed in clinical and industrial applications (Box 1).

Currently, the majority of phages and their host interactions remain unstudied in the laboratory, both because many host bacteria remain uncultivable and, as we will show, because techniques to isolate and study phages are biased[36]. For phages that have been tested, infection assays have shown that host-range spans a wide continuum, from extremely narrow to broad[163]. Recent studies of newly isolated marine broad host-range phage lineages (e.g.[164], [76]) and community-wide single-cell metagenomics efforts of extreme environments[55] indicate that broad host-range is potentially widely distributed in natural environments. This is supported by

a number of recent metagenomics efforts that have sought to link viral genomes to their hosts[82,119], suggesting a wide variety of host-ranges in nature (Figure 1). Such complementary and sometimes contrasting lines of evidence, combined with the importance of phage host-range and host-range variation for understanding phage biology, microbial ecology, and phage applications justify studying the specific molecular mechanisms that allow phages to infect multiple hosts.

Defining host-range

Like all aspects of phage biology, host-range can only be understood in the context of the phage life cycle. The best described are the lytic and temperate life cycles. Upon infection, obligately lytic phages rapidly produce offspring and lyse the host, while temperate phages may first integrate their genome into the host genome, until stressors prompt them to become lytic. Other life cycles, such as chronic infections and pseudolysogeny, are less widely described and their ubiquity remains unknown[165,166]. The lytic phase of the phage life cycle includes four stages. First, a phage encounters a suitable host, attaches to its surface, and inserts its genomic material into the host cell. Second, the phage bypasses host intracellular defences to establish a sustainable infection. Third, it replicates by hijacking resources of the host cell to produce progeny. Fourth, the phage progeny escapes from the host cell into the environment, restarting the cycle. Many studies into the molecular mechanisms of broad host-range phages focus on the attachment stage. For example, monovalent phages (i.e. binding to a single receptor) are more likely to have a narrow host-range, while polyvalent phages (i.e. binding multiple different receptors) may be able to infect more diverse hosts. However, each life cycle step involves circumventing defences and may require molecular and regulatory systems that are general enough to interact with multiple hosts. As we will discuss below, host-range modulation may depend on adaptations in the host receptor binding protein (RBP), but also on proteins that play a role in other life cycle stages. Together, we review current knowledge on the molecular mechanisms that allow broad host-range phages to infect multiple hosts and discuss the evolutionary context of the host-range phenotype.

Despite the importance of understanding host-range in the context of phage ecology and in potential phage applications, what constitutes broad versus narrow is not clearly delineated. The depth and breadth of bacterial and viral diversity, and difficulties in clearly delineating phage and bacterial taxonomy[167,168] make the formulation of specific definitions problematic. Also, both experimental and computational assessments of phage host-range have their

limitations[169], potentially leading to mislabelled host-range. Finally, phages are notoriously quick to mutate and evolve, and under the right circumstances phages may switch from narrow to broad host-range and back[170].

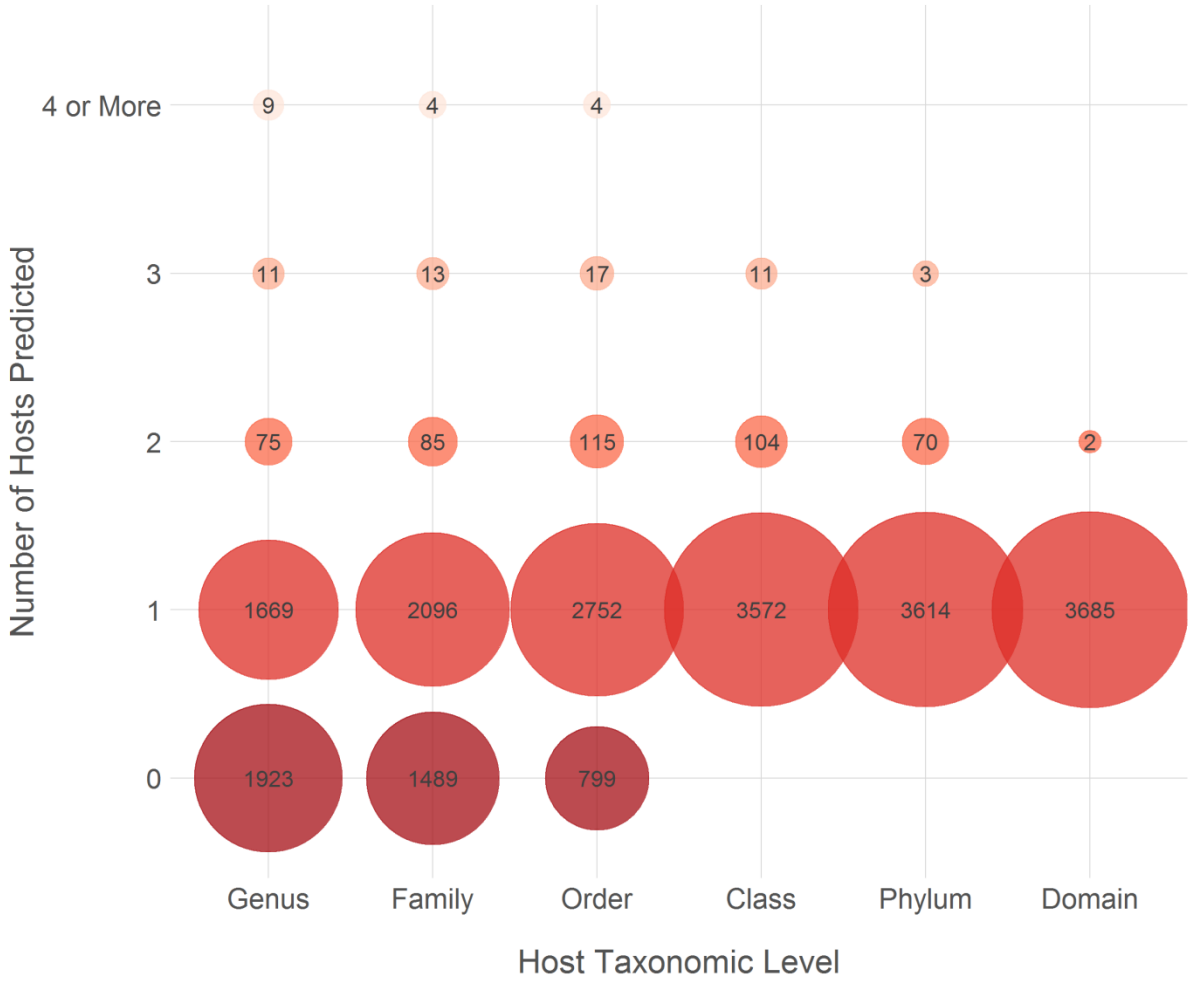


Figure 1. Meta-analysis shows diversity predicted host-ranges of phages identified by metagenomics. *A recent metagenomic virus discovery study[134] used various signals to predict the host-range of 15,280 detected viruses, including genetic homology, oligonucleotide usage patterns, clustered regularly interspaced short palindromic repeats (CRISPR) spacer matches, and tRNAs. Hosts were predicted for 3,687 out of 15,280 viruses. It remains difficult to make accurate host predictions, especially at lower host taxonomic levels: 1,764 hosts were predicted at the genus level and no species- or strain-level host predictions were made. Each column in the graph adds up to all 3,687 viral sequences for which at least one host prediction was made, for example 75 viruses had hosts predicted in two different genera, while 13 viruses had hosts predicted in three different orders, etcetera. Similar host-range patterns were observed in another recent study[119].*

For the purpose of this review, we will define narrow host-range phages as those that can only complete their life cycle in one host (Figure 2A), and broad host-range phages as those that can complete their life cycle in multiple taxonomically distinct hosts. There is no clear categorical distinction between different levels of broad host-range. Phages that infect multiple bacterial strains of the same species, as well as phages that infect bacteria from different genera are both said to have a broad host-range, although the latter is obviously broader than the former. Here, where relevant, we specify at which level the multiple infected hosts are taxonomically related.

We will further distinguish phenotypical and genotypical mechanisms facilitating broad host-range. The first involve individual broad host-range phage particles that are phenotypically capable of infecting multiple hosts (e.g. Phi92[171] or SP6[172], Figure 2B). The second are host switching phages where the entire phage quasispecies can infect multiple hosts, while individual particles only infect one host (Figure 2C). The broad host-range is thus the result of changes in the individual phage genotype. This difference may be important when formulating theoretical models to interpret microbial and viral dynamics when such data is available (Box 2).

In the following sections, we will first review the dedicated mechanisms that broad host-range phages possess at every stage in the life cycle, followed by a discussion of the experimental techniques that have been developed to isolate and study broad host-range phages. Together, we provide a comprehensive review of the current state of knowledge about phage host-range and hope to stimulate interest in this important topic.

Natural and engineered surface adhesion mechanisms of broad host-range phages

The initial interaction between phage and host is a chance encounter, in many studied phages followed by a two-step process of reversible and irreversible binding of the RBPs to host surface receptors[173]. As surface structures are among the most variable elements of a microbial cell and nearly every surface structure can serve as phage receptor[174], the initial stage of phage infection is an important determinant of phage host-range. Monovalent broad host-range phages may target structures on the host cell wall that are conserved between different hosts, such as some *Salmonella* phages that target the sugar cores of lipopolysaccharides in multiple hosts[175]. Polyvalent broad host-range phages encode mechanisms for binding multiple different receptors. The following will discuss the various phenotypic and genotypic mechanisms that

allow phages to bind to the cell walls of multiple hosts, as well as ways these mechanisms can be exploited to engineer broad host-range phages.

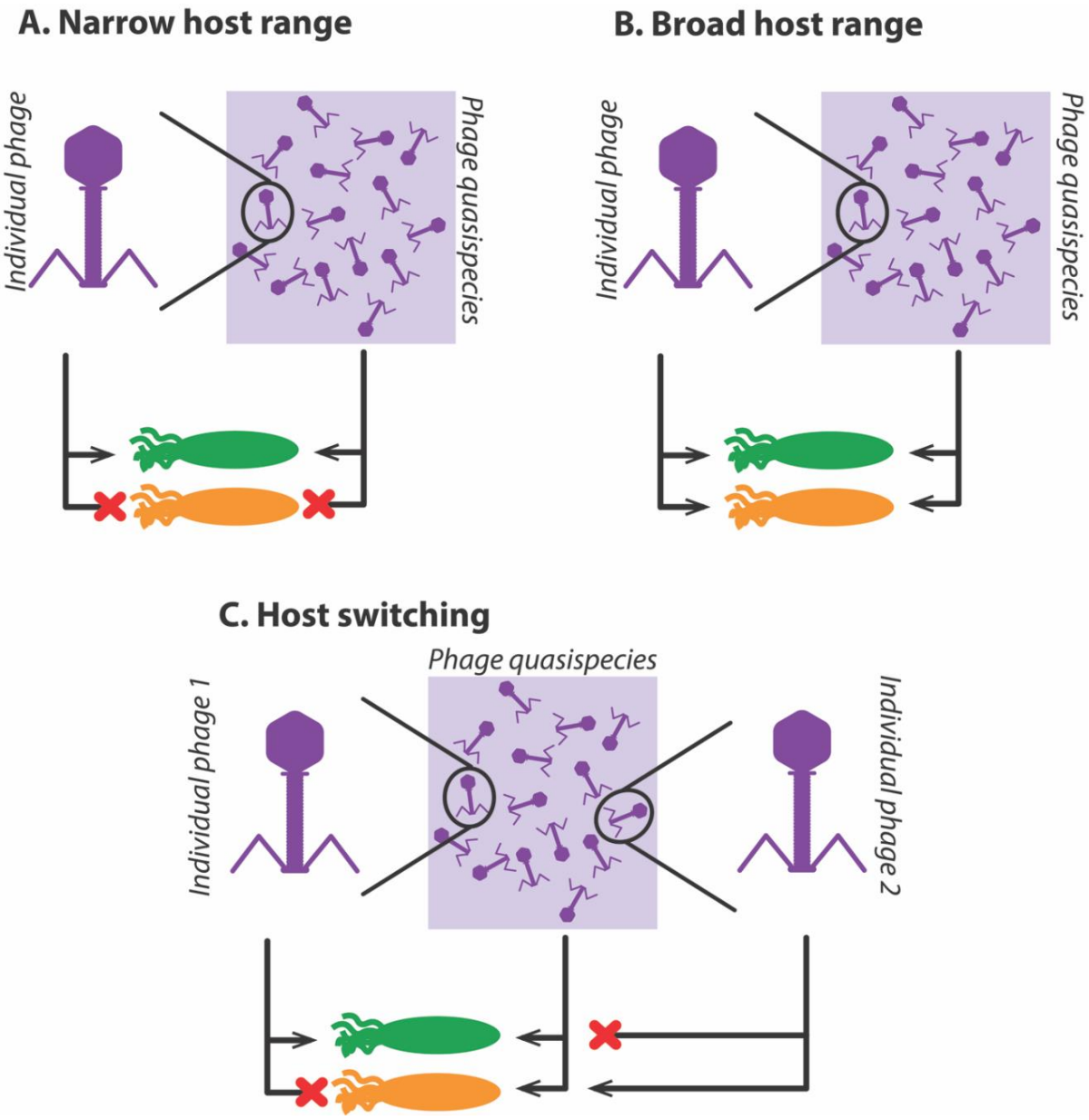


Figure 2. Subdivision of bacteriophages based on host range. *Phages can be divided into three basic groups based on their host range and the mechanism used to achieve it.* ***(A)*** *Narrow host range phages only infect one host (most studied phages).* ***(B)*** *Broad host range phage species infect multiple hosts, an ability present in every individual phage in the species (e.g. SP6[172]).* ***(C)*** *Host switching is a grey area between narrow and broad host range. Here, the quasispecies as a whole can infect multiple hosts, but each individual particle only has the capability to infect a single host (e.g. BPP-1[176]).*

Specific phage particles and a broad host-range quasispecies

Phage-host co-evolution studies have shown that host-specificity can be altered through mutations in the RBP gene[177–179] as well as other genes[180]. Phage host-range mutations can be caused by as little as one point mutation[181], this limited number allowing reversibility and repeatability in host-range evolution[178] (Figure 3A). However, most mutations do not lead to a change in host-range[182] due to the more constrained nature of gain-of-function mutations in the phage versus loss-of-function mutations in the host[177].

Polyvalent phage lineages with dual receptor specificity can result from RBP mutations. For example, the narrow host-range *Enterobacteria* phage λ can evolve into a broad host-range phage capable of infecting hosts from different genera by binding to either OmpF or the ancestral LamB protein receptors[178]. Further co-evolution led to specialisation of this broad host-range phage to either OmpF- or LamB-specific monovalent phages, even when both receptors were present in the environment[170]. Moreover, the genetically uniform broad host-range lineage also seemed to consist of two sub-populations preferring either LamB or OmpF, an effect hypothesized to be caused by differences in RBP folding[183]. These results show that the transition between monovalence and polyvalence can be readily made.

Some phages possess molecular mechanisms that allow for targeted genetic diversification of RBP, thereby allowing controlled host switching. In the *Bordetella* phage BPP-1[176], a reverse transcriptase allows for reverse transcription and adenosine-specific mutagenesis of a template region in its genome, which is then used to substitute the 3'-end hypervariable domain of the RBP-encoding *mtd* gene (Figure 3B). Another strategy was found in *Enterobacteria* phage T4, where duplication of hypervariable domains flanked by a GxHxH motif expanded its host-range from *Escherichia coli* to include *Yersinia pseudotuberculosis*, a different host within the order Enterobacterales[184] (Figure 3C). The discovery of such hypervariable regions in diverse phage contigs assembled from human faecal metagenomes suggests that the mechanism may be widespread[185]. For example, a similar system to the one in BPP-1 has been found in a taxonomically distinct intraterrestrial archaeal virus, ANMV-1[186].

Finally, some phages encode multiple RBPs on their genome while only expressing one at a time. *Enterobacteria* phages Mu and P1 encode two RBPs with distinct specificities. The orientation of a reversible genomic segment encoding both RBPs determines which RBP is expressed and therewith the specificity of the phage particles (Figure 3D)[187].

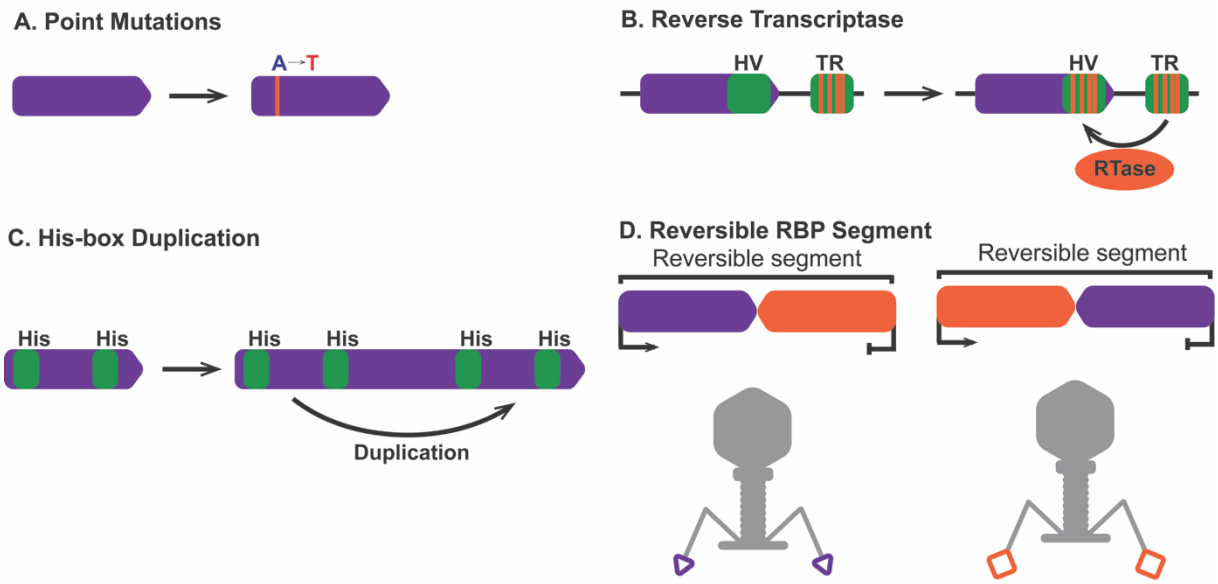


Figure 3. Mutational strategies for broadening or switching host-range. *Some phages are known to employ methods to mutate RBP to change receptor specificity. While each individual virion is monovalent, these phages represent a polyvalent quasispecies.* ***(A)*** *Point mutations in the gene encoding the receptor binding protein (RBP) may alter the structure of the protein. These point mutations are random, although convergent evolution has been observed (examples in main text).* ***(B)*** *The polyvalent Bordetella phage BPP-1 alters its host specificity through a reverse transcriptase mediated mechanism. This protein introduces a combination of specific mutations to a template region (TR) upstream of the RBP. This subsequently replaces the hypervariable region (HV) at the 3' end of the gene*[176]*.* ***(C)*** *In phage T4, variable sections of the RBPs are flanked by GxHxH motifs called His-boxes. Duplication of these His-boxes changes RBP specificity*[184]*.* ***(D)*** *Phages Mu and P1 have two RBPs with different specificities organised in a reversible genome segment. This segment can be flipped around in the prophage state, thereby determining which RBP is transcribed*[187]*.*

Broad host-range phage particles

Recent cryo-electron tomography (cryo-ET) studies have revealed polyvalent phages that simultaneously express multiple RBPs. *Salmonella* phage SP6 can infect multiple *S. enterica* subsp. *enterica* serovars by expressing two RBPs on a V-shaped structure[172] that swivels to change tailspike orientation depending on the host species encountered (Figure 4A). Similarly, *Escherichia coli* phages DT57C and DT571/2 encode dual RBPs that are hypothesized to branch out from the phage tail fibre, possibly in similar fashion to SP6 RBPs[188]. This mechanism also occurs in phages that infect Gram-positive bacteria, like *Staphylococcus* phage φSA012, which uses a main RBP to bind to the teichoic acid core of some *Staphylococcus*

aureus strains and a second RBP to bind an *α*-N-acetylglucosamine moiety in *S. aureus* strains where the teichoic acid core is unavailable due to glycosylation[175].

Enterobacteria phage Φ92 encodes no fewer than five RBP genes for infection of several different *E. coli* and *Salmonella* strains[171]. Cryo-ET showed at least three distinct RBPs radiating from the Φ92 baseplate, each in six-fold (Figure 4B). It remains unknown if the other two encoded RBP genes are functional or if different selections of the five RBPs are expressed at a time. These results show how recent developments in electron microscopy now allow the study of phage-host interactions in greater detail than before. Although only the SP6 and Φ92 mechanisms have been studied in detail, future studies will undoubtedly uncover other broad host-range mechanisms in phages that encode multiple putative RBPs, like *Sinorhizobium* phages ΦM9 and P10VF[189].

An intriguing alternative for phages to overcome receptor specificity without encoding multiple RBPs has been observed in *Bacillus subtilis* cultures infected with *Bacillus* phage SPP1[190]. Phage-induced cell lysis promotes formation of membrane vesicles[191], some of which contain the SPP1 receptor. When such vesicles merge with resistant cells, they may become transiently sensitive to the phage. While this might be a rare chance event, it could effectively increase the host-range of monovalent phages.

Bioengineering phage receptor binding proteins

Bioengineering RBPs is an important potential tool to control phage host-range in biomedical applications. Broader host-range has been engineered through recombination of homologous RBPs, whose C-terminus often contains the receptor binding domain. For instance, recombining the RBP C-termini of *Enterobacteria* phages T3 and T7 expanded the host-range of each beyond the sum host-range of both[192]. Similarly, recombination of RBP domains from the T4-like *Enterobacteria* phages WG01 and QL01 resulted in phage WQD, whose host-range includes that of both parent phages as well as several novel hosts[193]. A firm understanding of RBP structure and functioning is thus needed for future engineering efforts. Nevertheless, the fact that merely exchanging the tail fibre of phage T3 with that of *Yersinia* phage R resulted in T3 gaining the ability to infect the phage R host *Yersinia tuberculosis* shows the potential of bioengineering the RBP for host-range expansion[192]. In addition to increasing the host-range of phages through bioengineered RBPs, plasmids containing foreign RBPs can be spread throughout bacterial populations by transducing phages like T7[194].

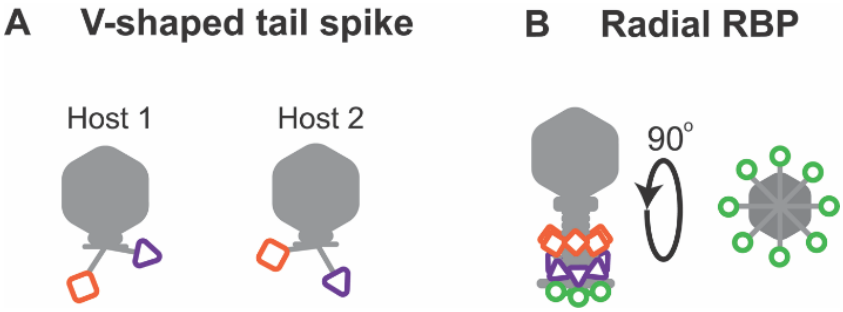


Figure 4. Phages encoding multiple receptor binding proteins with differing specificity. *Some phages encode and/or express multiple RBPs to allow binding to multiple receptors, thus representing polyvalent virions.* ***(A)*** *Phage SP6 possesses two RBPs at the end of a V-shaped tailspike. Changing the angle of the tailspike determines which RBP can bind to a receptor*[172]. ***(B)*** *Phage phi92 encodes at least five RBPs in its genome. At least three of these are expressed at the end of the tail fibre. It is unknown whether the remaining two RBPs are expressed*[171].

Aspects of host-range after entry into the host cell

Genome replication

After entry into the cell, several factors influence phage host-range. Phages need adaptations to the molecular mechanisms of the host (e.g. codon usage) and evade intracellular defence systems, like restriction modification (RM) systems, clustered regularly interspaced short palindromic repeats (CRISPR), abortive infection (abi), bacteriophage exclusion (BREX) and defence island system associated with restriction–modification (DISARM)[195–197]. Phages have developed myriad ways to counter specific intracellular defences, including anti-CRISPR proteins[198], RM inactivation proteins[199], and the recently discovered phage nucleus that protects *Pseudomonas* phage 201φ2-1 DNA during replication and transcription[200]. General anti-defence mechanisms such as the latter phage nucleus might be useful assets to broad host-range phages. With the ongoing characterization of unknown phage proteins[201], we expect that novel mechanisms to overcome host intracellular defences will be uncovered.

Akin to the diversity in bacterial outer membrane structures, bacteria can defend against phages by encoding unique transcriptional responses to infection[202]. While broad host-range phages might thus benefit form specialising transcriptional programs to each host, the opposite is true in phages like *Synechococcus* phage Syn9 that infects different environmental isolates, as it showed a uniform transcriptional program despite the different responses in three hosts[202]. The same holds for the broad host-range *Bacteroidetes* phage φ38:1 that showed a uniform transcriptional program in response to varied transcriptional programs in closely related

Bacteroidetes isolates, leading to differences in infection efficiency in those hosts[203]. Phage φ38:1 successfully overcomes intracellular defences and transcriptional responses in its original isolation host, but not in other hosts[204]. These results could indicate that it is difficult for broad host-range phages to evolve specialised transcriptional programs for different hosts. Still, the conserved aspects of phage transcriptional programs[205], might be selected for in broad host-range phages.

While phage host-range is often discussed in the context of lytic phages, it is a similarly relevant characteristic of temperate phages. The unique challenge for temperate broad host-range phages is the integrase-mediated integration of their genome into that of their hosts. Integrases rely on distinct target sequences to perform this integration, e.g. within conserved tRNA genes[206]. Thus, the process is often host-specific[92], or more precisely it is limited to hosts with the correct integration target site. *Brucella* phage BiBPO1 is capable of integrating in the genomes of a broad range of hosts by targeting conserved integration sites in tRNA genes[207]. An alternative temperate lifestyle is to circumvent integration altogether by maintaining the genome as a plasmid[208], as was observed in the temperate broad host-range *Ochrobactrum* phage POI1126[209]. Interestingly, the widespread crAssphage does not encode an integrase protein and has never been found integrated in any bacterial genome, but it does have two proteins with putative plasmid replication domains on its genome[83,86]. It may thus employ this mechanism to infect diverse hosts, potentially contributing to its widespread abundance.

Exiting the host cell

To complete the lytic life cycle, a broad host-range phage needs to lyse potentially divergent cell wall structures and release its progeny. Many phages carry an endolysin which lyses the cells, and a holin which allows the endolysin access to the periplasm at a specific timing in the infection[210]. Endolysins can have lytic host-ranges that are broader than the natural host-range of the phage[211] and may thus not form critical bottlenecks in the *in vivo* lytic range. However, the host-range may also be bounded by the ability of phage holins to give endolysins access to the cell wall. Moreover, bacteria may suppress phage endolysins to prevent phage progeny from leaving the cell[212], so a broad host-range phage still needs to possess an altogether compatible lysis system to efficiently exit the various host cells.

Studying phage host-range

As the previous sections show, multiple phages with broad host-range and their molecular mechanisms have been identified, and ecological as well as metagenomics studies indicate that broad host-range may be more prevalent than previously thought. Yet, narrow host-range is prevalent among isolated and well-studied phages. Understanding this disparity demands insight into the evolutionary dynamics of host-range (Box 2) and the methodological biases inherent in phage isolation techniques.

The ecology of phage host-range

In phage-host co-evolutionary experiments, an increase in the host-range of a phage often coincides with a decrease in the virulence of that phage on each individual host. This trade-off reflects antagonistic pleiotropy, where a mutation is beneficial to one trait but detrimental to a second. Experimental evolutionary studies, for example with *Pseudomonas* phage φ6[213], have shown evidence of pleiotropic costs to host-range expansion. Examples where such trade-offs are evident include: (i) a recent study of the non-tailed *Autolykiviridae* family of marine *Vibrio* phages that have broader host-ranges, but are slower growers than their tailed counterparts[76], (ii) a trade-off between host-range and growth rate when comparing *Enterobacteria* phages with different host-ranges within the *E.coli* ECOR strain collection[214], and (iii) a trade-off between host-range and thermostability in polyvalent phage λ strains[183]. Antagonistic pleiotropy raises a substantial evolutionary barrier to host-range expansion in the environment, but it is not universal in phage host-range evolution. It has been proposed that in complex phage-host systems, hosts that evolve resistance to one phage can become susceptible to another phage[215]. For the latter phages, this means that they passively gain infectivity of a new host without incurring mutations or pleiotropic costs[182].

Host density, diversity and quality are likely relevant factors in determining phage host-range, as described in optimal foraging theory[216]. This theory predicts that host-range expansion is only favoured if the potential novel host is either present at higher densities or is of higher quality than the original host. This is illustrated by the example of *Pseudomonas* phage φ2, that only expanded its host-range if the density of susceptible hosts was less than 1% that of resistant hosts[217]. While optimal foraging phage studies successfully predicted the outcome of competition between narrow and broad host-range phages[218], studies with *Pseudomonas* phage φ6 showed that the evolution of generalists depended on both host quality and competition

intensity[219]. This may mean that broad host-range is selected for in oligotrophic environments, where cell densities are low but species diversity is high[164].

Biases in experimental approaches

Several phage isolation techniques use mono-cultures with a high host density, either in liquid or double-layer agar medium. These popular assays can only detect phages that are able to complete a full lytic cycle in the given experimental conditions and within the time frame of the experiment. Moreover, pleiotropic trade-offs between host-range and virulence, discussed above, may result in broad host-range phages with slower growth rates being outcompeted by narrow host-range phages with higher virulence under these conditions[76,218]. Beside these ecological biases, many methods including double-layer agar and liquid assays only test a single phage-host association at a time, so perceived phage host-ranges can only extend insofar as the hosts that it was tested upon. Methods such as density gradients, chloroform extractions, and DNA isolations have been shown to contain biases for tailed dsDNA *Caudovirales* that make up >90% of studied phages[76]. Studies have shown that non-tailed phage lineages are dominant in some marine environments[220,221], and in the case of non-tailed *Autolykiviridae*, have been found to have a tendency for broad host-ranges[76]. Finally, it has been noted that some methods, like spot assays, may overestimate phage host-range, meaning evaluations of phage host-range must be treated with some caution[222].

Methods are increasingly available that circumvent isolation bias for narrow host-range phages[223–225]. In particular, methods are being developed to isolate broad host-range phages infecting different genera. Co-cultures of two potential hosts allowed the isolation of phage SN-T, that formed plaques on hosts from six genera in spot assays[224]. Similarly, sequentially culturing on monocultures of multiple potential hosts resulted in phages that form plaques on multiple strains in the genus *Pseudomonas*, as well as *E. coli*, a different host from the class *Gammaproteobacteria*[225]. Culture-independent techniques, as reviewed in[92] avoid some of these biases, although they come with other biases of their own. Examples include microfluidic PCR[226], phage fluorescent *in-situ* hybridization (FISH)[227], and metagenomic chromosome confirmation capture (Meta3C)[228].

Biases in computational approaches

Computational methods to determine phage host-range also have biases. While accurate, identification of matching CRISPR-Cas spacers is limited to detecting host-associations of

microbes that contain an active CRISPR system[92]. Another popular technique is detecting similarity in oligonucleotide usage profiles between the genomes of phages and of potential hosts, but this similarity score is a sliding scale, and reliability thresholds remain obscure[134]. Machine learning techniques hold promise but urgently require diverse and high-quality training data in the form of high-throughput phage-host interaction screens of naturally occurring viruses. The currently known phage-host interactions available in databases remain a limited and as explained above, a biased subset of the naturally occurring interactions. Due to the sensitivity of machine learning techniques to training data biases, these techniques should be treated with caution in the face of the diversity of the natural virosphere.

Concluding remarks

Three major developments have catalysed the surging interest in bacteriophage research in recent years. First, viruses observed in environmental samples through epifluorescence microscopy[39] and metagenomic screens consistently showed that the majority of natural viruses remain uncharacterized[36,83]. Second, discoveries including CRISPR-Cas have led to biotechnological innovations[93], adding interest in previously unknown aspects of phage-host interactions. Third, phage therapy seems a promising weapon to combat the global rise of antibiotic resistance in bacterial pathogens[229].

Studying the molecular mechanisms of phage host-range and specificity is essential for a wide variety of clinical and industrial applications. We have shown that specificity is a complex notion, dependent on many mechanisms at every step of the phage life cycle. As we continue to find novel phages and study their host interactions, current notions on host-range will continue to adapt. The ongoing co-evolution between microbes and their viruses has driven a multitude of innovations and thus novel fascinating discoveries with paradigm-shifting consequences are to be expected (listed in the Outstanding questions section). Indeed, the increased interest in phage biology has recently led to discoveries of remarkable molecular mechanisms in phages, including phages forming nucleus and tubulin-like structures upon infection[200], phages incorporating eukaryotic genes into their genome[230] and phages engaging in quorum sensing [20].

Meanwhile, advanced computational models of phages and their hosts including their eco-evolutionary dynamics are needed to integrate our current knowledge on microbial ecosystems and reveal gaps in our understanding[231,232]. Data from recent large-scale environmental surveys, including viral metagenomic sequencing and genome assembly will

enable the integration of phages into ecological models to study the global importance of microbes and their phages, ranging from nutrient cycling to gene flow at scales ranging from the global oceans to the human gut.

BOX 1: Host-range in phage applications

Several applications critically depend on the host-specificity of phages. Phage typing utilizes the specific nature of phage-host interactions by generating a phage susceptibility pattern unique to a singular bacterial pathogen [233]. Such patterns can subsequently be used to distinguish between bacteria and track different pathogenic strains. As identifying pathogens using phage typing is wholly dependent on the specificity of the lysis pattern, phages with highly stable host affinity are essential.

With the rise of antibiotics resistance among pathogens worldwide, phage therapy is increasingly considered as a viable alternative for application both in medicine and in the food industry[229,234]. Although the idea behind phage therapy is as old as phage biology, recent years have seen an impressive revival (e.g. [235,236]). In phage therapy, phage host-specificity is a double-edged sword. On the one hand, narrow host-range phages preserve the natural microbial flora and lower the probability of developing community-wide resistance[237,238]. On the other hand, such phages may need to be isolated for each target strain, increasing labour requirements for their isolation and validation. Moreover, pathogens may readily evolve resistance to individual phages, so infections need to be treated with phage cocktails where each of the members are to be validated[239]. Broad host-range phages may be a promising solution, but before this can be developed the evolutionary stability of phage-host interactions needs to be established.

Efforts to control microbial community composition further hinge on an accurate understanding of phage host-range and its evolution. For example, temperate phages have been used to spread sensitivity to antibiotics in *E. coli*[240,241] and to deliver a CRISPR-Cas9 system designed to specifically kill pathogenic bacteria[242]. In such applications, broad host-range was found to be an asset when targeting specific bacteria in heterogeneous environments[161].

BOX 2: Eco-evolutionary models of phage host-range

The main tools to study the ecological and evolutionary significance of the host-range continuum include theoretical models[243], statistical analyses[244], and co-evolutionary experiments[245]. These studies have led to two major competing models of phage-host co-evolution: arms race dynamics (ARD) and fluctuating state dynamics (FSD)[246].

In ARD, host and phage iteratively evolve new resistance and infectivity, respectively, while maintaining resistance and infectivity to ancestral antagonists. Thus, over time, phages evolve a broader host-range which includes progressively more resistant hosts. When depicted in a matrix, the resulting interaction network is nested[244], where the more specialist phages iteratively infect subsets of the hosts relative to the more generalist phages. Antagonistic pleiotropy of infective traits in phages would keep phages with universal host-ranges from evolving in such a system [247].

In FSD, phage-host interactions remain highly specific. When hosts evolve novel resistance, phages overcome this to maintain infectivity, but in the process lose the ability to infect ancestral hosts. In a similar fashion, hosts are only resistant against co-occurring phage strains but not to ancestral strains. This causes a fluctuating selection between highly specific host and phage genotypes, an extreme case being a monogamous network of phage-host interactions where one phage infects one host[243]. However, modular networks may be more likely in heterogeneous natural environments, where phages and hosts interact in subsets.

Small scale, local networks often show nested structures[244], while modularity becomes evident at larger scales[248,249]. Host-range is likely bounded by the phylogenetic, geographical, and ecological distance between a phage and its potential hosts[250], phages not always being adapted to local hosts[251]. Phages have been observed to be specific not so much to a species but to a locale[245,252], with different locales containing different frequencies of broad host-range phages[253]. Finally, interaction networks imperfectly reflect co-evolutionary dynamics[254], so it is clear that many mechanisms drive phage evolution[182] and much remains to be discovered.

4

Adsorption sequencing as a rapid method to link environmental bacteriophages to hosts

"Go back?" he thought. "No good at all! Go sideways? Impossible!
Go forward? Only thing to do! On we go!"
The Hobbit, J.R.R. Tolkien

Published as: de Jonge, P.A.; von Meijenfeldt, F.A.B.; Costa, A.R.; Nobrega, F.L..; Brouns, S.J.J.; Dutilh, B.E.; *Adsorption sequencing as a rapid method to link environmental bacteriophages to hosts;* iScience; 2020; 23; 101439.

Summary

An important challenge of viral metagenomics studies is to associate bacteriophages to hosts. To address this challenge, we developed adsorption sequencing (AdsorpSeq), a readily implementable method to measure the key initial step in the infection cycle, the adsorption of phages to host cell surface. In AdsorpSeq, phages are added to bacterial cell envelope suspensions. Phages that adsorb to the cell envelopes are separated from free phages through agarose gel electrophoresis, after which DNA from adsorbed phages is sequenced. Here, we show that AdsorpSeq allows for the separation of model phages based on receptor-adsorbing capabilities. Next, we applied AdsorpSeq to identify phages in a hospital wastewater virome that adsorb to the cell envelopes of nine taxonomically distinct bacteria, including important pathogens. We detected 26 adsorbed phages including common and rare members of the virome, a minority being related to previously characterised phages. We conclude that AdsorpSeq is an effective new tool for rapid characterisation of environmental phage adsorption to host cell envelopes.

Introduction

Bacteriophages (viruses that infect bacteria) are omnipresent and impact every ecosystem[36]. Their impact on microbial communities makes phages both useful and detrimental. On the one hand, they are potential bioengineered drug delivery systems[255] and alternatives to antimicrobials[238]. On the other hand, they spread bacterial pathogenicity[256] and disrupt food production chains like milk fermentations used in the dairy industry[257]. Phages also affect ecosystems at larger scales by controlling bacterial evolution and community structure, affecting e.g. our microbiomes[63,65], marine nutrient cycling through bacterial lysis[56,258], and global oxygen production by encoding photosynthesis genes that are expressed in cyanobacterial hosts[259]. While their global importance makes understanding phage-host interactions crucial, most remain undetermined [36].

A major reason for the mass of undetermined phage-host interactions is a shortage of readily applicable experimental techniques that can simultaneously 1) identify phages in an environmental sample, and 2) link them to host(s). Unstudied phage genomes can be identified with metagenomics techniques, but despite constant improvements (e.g.[127,260–264]) it remains challenging to predict to which hosts these phages adsorb, especially at low taxonomic levels[92]. While CRISPR-Cas memory (spacers) and prophage regions can result in reliable host-

predictions[92], most bacteria have no CRISPR-Cas systems[94] and not all phages form prophages. Beyond computational approaches, phages can be linked to their host with isolation techniques like double-layer agar plates. Such techniques depend on phages to form visible plaques[81], and can be biased to phages with narrow host ranges[76,218]. These assays furthermore often employ a few highly related hosts, each needing a separate assay[222]. Thus, available information on phage host-range is limited. Other proposed methods include meta3C, which infers interactions based on physical proximity of phage and host DNA[228] , and those summarised in an earlier review[92]. However, such methods are generally cumbersome and thereby hard to implement.
An alternative approach to determining phage-host interactions is by focusing on the first step of the phage infection cycle, the adsorption of the phage to bacterial surface receptors. While phage adsorption is not always followed by successful phage infection, it is a crucial step for successful infections and often specific[128]. Utilizing phage adsorption specificity could thus allow studies of phage-host interactions in environmental samples. This was recently shown through viral tagging[88,265], where fluorescently labelled phages are added to bacteria. Bacteria bound by fluorescently labelled phages are isolated with fluorescence-activated cell sorting, and phage-bacterium pairs are isolated. This approach allows abundant viruses to be linked to hosts, but it can remain challenging to identify phage-host links for rare members of the virome. Finally, viral tagging requires a highly specialised experimental setup.

Here, we rapidly identify phage-host pairings by linking cell envelope-adsorption to phage sequencing and statistical analysis (adsorption sequencing or AdsorpSeq). AdsorpSeq allows identification of novel phages and their host interactions by exploiting differential migration of phages bound to host receptors and unbound phages in agarose gel electrophoresis. This enables selective sequencing of phages based on their interaction with cell envelopes of a specific host. Thereby multiple phages that interact with a given host can be rapidly and simultaneously identified. We show that model phages can be differentially identified based on the presence of their receptor molecule. Subsequently, we apply AdsorpSeq on a hospital wastewater virome and the cell envelopes of nine taxonomically distinct bacteria, uncovering 26 novel phage-host interactions.

Results and Discussion

Identification of model phages based on adsorption to their hosts

AdsorpSeq aims to selectively sequence phages based on their adsorption to bacterial cell envelopes. This is achieved by a five-step process (steps 1-5 in Figure 1A). First, a phage

mixture is added to a cell envelope suspension that was isolated from a bacterium of interest (step 1 in Figure 1A). An incubation then allows phage adsorption to their receptors (2). Next, agarose gel electrophoresis separates bound and unbound phages (3). Unlike unbound phages, phages bound to cell envelope suspensions will migrate slower into agarose gels due to the larger size and altered charge of the adsorption complex. The result is rapid separation of phages based on adsorption abilities. Finally, genomic material of bound phages is isolated from the gel (4) and sequenced (5).

To validate AdsorpSeq, we tested the method with *Escherichia* phage λ and *Salmonella* phage P22 as two model phages with well described adsorption properties. These two phages differ in host, morphology, and the receptor type. The receptor of *Siphoviridae* phage λ is the *E. coli* maltose pore protein LamB[266], while the *S. enterica* subspecies *enterica* (hereafter: *S. enterica*) lipopolysaccharide O-antigen chain serves as receptor for *Podoviridae* phage P22[267]. Because absorption to the bacterial cell envelope does not guarantee successful infection[128] AdsorpSeq may detect phage-bacteria interactions beyond infecting host range. However, as both phages are specific to their respective receptors[267,268], we could gauge preservation of adsorption specificity in AdsorpSeq.

As a first test, we added either phage λ or P22 to *E. coli* and *S. enterica* cell envelope suspensions. Upon adding non-host cell envelopes (e.g. adding *S. enterica* to λ), phage particles migrated into agarose gels, while adding host cell envelopes (e.g. adding *E. coli* to λ) resulted in phage particle retention around the sample slot at the origin of electrophoresis (Figure 1B). These gel regions consequently contained significantly more DNA when using host cell envelopes than when using non-host cell envelopes (two-tailed t-test, $p < 0.05$). This showed that adsorption of phage particles to host cell envelopes prevented migration into agarose gels. To confirm receptor specificity, we repeated the experiment with a LamB-knockout strain (ΔLamB)[269], which resulted in phage λ losing adsorption ability (Figure 1C). These results agree with earlier studies[267] showing slower migration of phage particles into agarose gels after addition of purified phage receptor particles.

After establishing that presence of host cell envelopes alters phage migration in agarose gels, we used this property to preferentially sequence phage genomes from a mixture based on the presence of host cell envelopes. We performed AdsorpSeq with a mixture of equal parts phage λ and P22, which we added to cell envelopes of either *E. coli* or *S. enterica*. Upon sequencing of genomic material isolated from agarose slices, two to three times more reads mapped to the phage λ genome than the P22 genome when *E. coli* cell envelopes were added and vice versa (Figure 1D). The resulting difference in phage genome abundance was highly

significant (Fisher's exact test, $p = 2.5 \cdot 10^{-16}$). As AdsorpSeq was thus capable of discerning phage-host associations in a simple phage mixture, we next applied it to a complex environmental phage mixture.

AdsorpSeq results in selection of unique phage subsets

We next applied AdsorpSeq to identify phages targeting specific bacterial cell envelopes in a complex virome derived from a hospital wastewater influent pipe (Supplementary Figure 1A). Phage adsorption targets consisted of cell envelope suspensions from nine taxonomically diverse bacteria, including three *Enterobacterales* (*Escherichia coli, Citrobacter freundii,* and *Klebsiella pneumoniae*), two *Pseudomonadales* (*Pseudomonas aeruginosa* and *Acinetobacter baumannii*), two *Bacteroidales* (*Bacteroides fragilis* and *Bacteroides dorei*), one *Burkholderiales* (*Ralstonia pickettii*), and one *Fusobacterales* (*Fusobacterium necrophorum*). All these bacteria are either part of the healthy human gut microbiome (e.g. *B. dorei*)[64] or pathogens linked with hospital infections (e.g. *K. pneumoniae*)[270]. Next to the virome treated with the cell envelopes of these nine bacteria, we sequenced the full untreated virome as control. As we increased DNA quantities of all samples by multiple displacement amplification (MDA), which alters apparent viral community compositions[271], the full virome was sequenced both before and after multiple displacement amplification (MDA). This allowed us to gauge and correct for MDA effects during data analysis (below).

Sequencing of the nine cell envelope-treated samples and the two viromes (pre- and post-MDA) resulted in 138 Mbp of read-level data. A cross-assembly resulted in 23,373 contigs longer than 2,500 bp, representing 71.4% of the total dataset, as determined by mapping the reads back to the contigs (for annotated contig metadata and contig abundances, see Supplementary Table 1). Taxonomic classification showed that the cross-assembly contained 1,111 viral contigs and a further 8,921 contigs that were taxonomically unclassified, while the remaining 13,341 contigs were mostly derived from bacteria, with a minority of archaeal and eukaryote contigs. The group of unclassified contigs likely reflects the large numbers of unstudied human gut phages[60]. We therefore combined the viral and unclassified sets to create a dataset of 10,032 confirmed and suspected viral contigs, which represented 51.5% of total reads. The percentage of reads represented by selected contigs fluctuated across the cell envelope-treated samples, ranging from 41.0% (*E. coli*-treated) to 81.9% (*B. fragilis*-treated) of reads (median: 70.2%, Supplementary Figure 1B). These differences between the samples were a first indication that phage adsorption depended on the cell envelope suspension used.

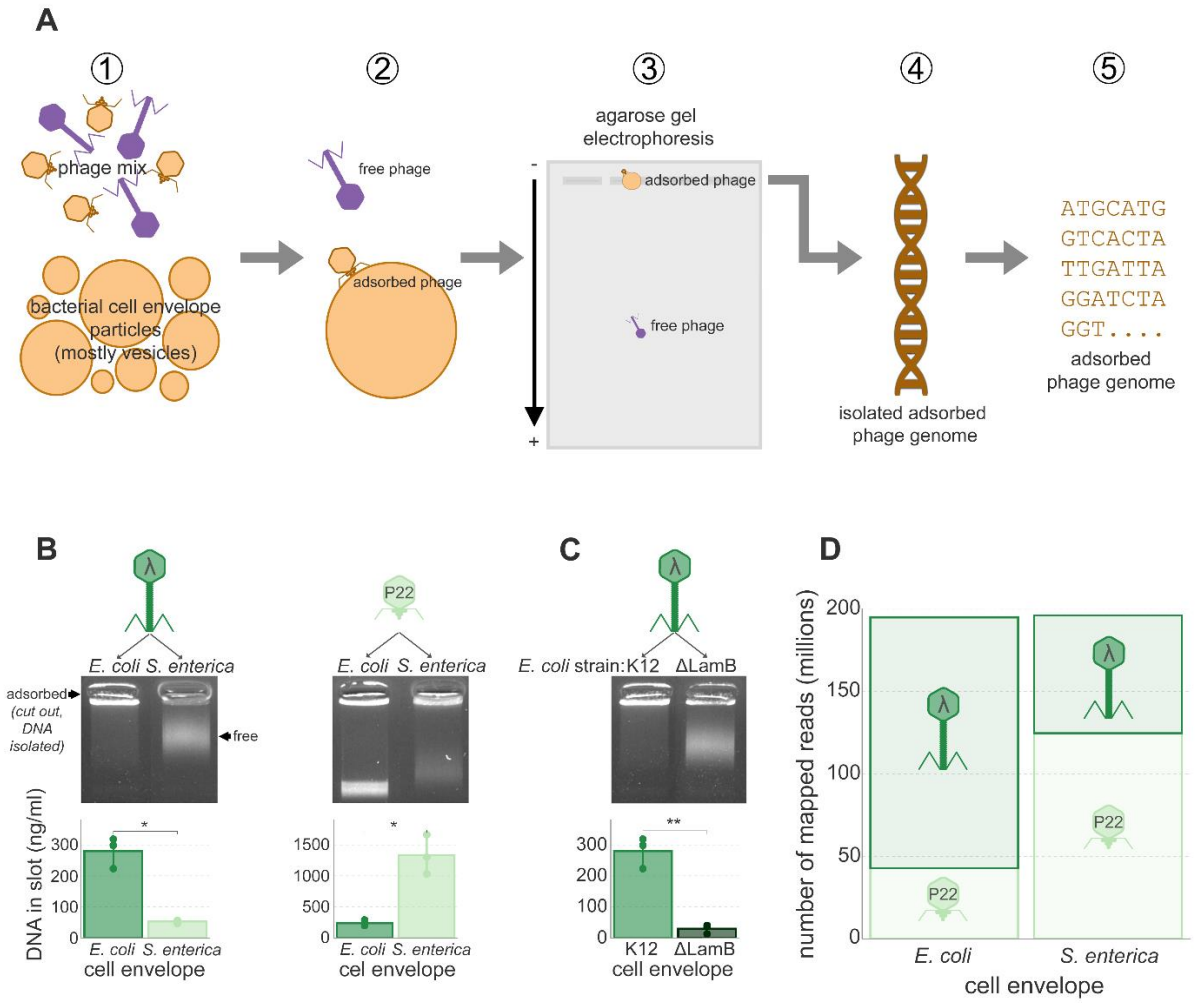


Figure 3: AdsorpSeq allows the selective sequencing of model phages based on adsorption. ***(A)*** *Schematic of AdsorpSeq. It shows the main steps of (1) mixing phages with bacterial cell envelopes, (2) allowing phages to adsorb to cell envelopes, (3) separating phages using agarose gel electrophoresis based on adsorbing capability, (4) isolating the genomes of adsorbed phages, and (5) sequencing genomes of adsorbed phages isolated from gels.* ***(B)*** *Adsorption of phages λ and P22 to host cell envelopes hinders their migration into agarose gels. Agarose gels of phages λ and P22 after being added to cell envelope suspensions of E. coli K12 and S. enterica S1400, and bar graphs showing DNA quantities that were isolated from the gel slots at the top of the gels. Arrows indicate the location of free phages (migrated into the gel) and adsorbed phages (in the gel slot) in the first gel. This is identical in the other gels.* ***(C)*** *AdsorpSeq maintains receptor molecule specificity of phage λ. Agarose gel of phage λ after being added to E. coli strain K12, to which it can adsorb, and E. coli ΔLamB, to which it cannot adsorb. Bar graph depicts DNA isolated from gel slots at the top of the gel. Note: although the smear seems visually stronger in the K12 lane, significantly more DNA was retained in the well containing the K12 envelope fraction than in the ΔLamB envelope fraction (see bar graphs).*

(D) *Applying AdsorpSeq to a mixture of phages leads to differentiation based on adsorbing capacity. Stacked bar graph showing the number of reads mapped to phages λ and P22 after AdsorpSeq was applied using an equal mixture of the two phages and cell envelopes from either E. coli K12 or S. enterica S1400. Significance levels according to a paired t-test, error bars depict standard deviations, points are biological replicates. *p<0.05, **p<0.01.*

While there were 842 contigs with identical ends in the selected dataset representing putatively complete and circular genomes, many contigs also represented likely genomic fragments. We therefore assigned contigs with similar tetranucleotide usage patterns and read depth patterns across the nine cell envelope-treated samples to 1,058 viral populations. Jaccard distances based on viral population relative abundances showed distinct dissimilarity between the cell envelope-treated samples (Supplementary Figure 1C). This indicated that each sample contained a unique set of viral populations and thus selected different phages. The viral populations with the highest abundance across the samples were also highly abundant in both virome controls (Supplementary Figure 2A and 2B) suggesting that some phage particles are retained in the wells through non-specific interactions, as also observed for λ and P22 (Figure 1D). Perhaps some phages are physically prevented from migrating into agarose gels by the vesicles that cell envelope suspensions likely form[272]. While we did not achieve perfect separation between bound and unbound phage particles, we concluded from the evident dissimilarity in composition between the samples (Supplementary Figure 1C) that selection of phages by AdsorpSeq is dependent on the type of bacterial cell envelope used.

Selection of phages with putative adsorption activity

To identify viral populations that specifically adsorb to the cell envelopes of one of the one of the nine bacterial species, we selected viral populations that were overrepresented one or more sample. To focus on the strongest adsorbing phages first, we defined overrepresented viral populations based on outlier analysis (see Methods for details), which meant that we selected viral populations with a relative abundance of at least 1.58 times higher in one sample than in the other eight samples. Relative abundance values from samples of the same bacterial taxonomic order were discounted when determining overrepresented viral populations to allow for phages with a broad adsorption or infection host range. In total, 123 viral populations represented phages that specifically adsorb to the cell envelope fractions of one of the nine bacterial species (Supplementary Figure 3A).

We next refined the selection of putatively adsorbing viral populations by applying two filters. The first filter removed viral populations that were positively selected for by MDA. MDA can result in efficient rolling circle amplification, but has also been shown to lead to a bias for small ssDNA phage genomes[273–275]. This held true in our dataset, as comparing viral populations in the virome before and after MDA showed 10-100-fold higher amplification of ssDNA *Microviridae* than other phage families (Supplementary Figure 3B). To reduce the impact of this bias, we thus filtered out 79 viral populations with strong MDA selection, leaving 44 viral populations that passed the MDA selection filter (Supplementary Figure 3A).

In addition to MDA selection, we tested if certain phages were universally selected for by the AdsorpSeq technique (Supplementary Figure 3C). This identified 18 putative adsorbing viral populations for which the relative abundance in all nine samples was higher than in post-MDA virome (Supplementary Figure 3A). This methodological bias was highest among *Inoviridae* and *Microviridae* (Supplementary Figure 3C). Although *in silico* evidence suggests that some phages may have very broad host ranges[119,134], most phages likely have a narrow host range spanning a few closely related strains within the same species or genus[88,128]. We thus interpreted our findings as a methodological selection bias that may reflect the inability of a phage particle to migrate into agarose gels due to large size, low charge, or non-specific interactions with bacterial cell envelopes. This may explain the stronger methodological selection pressure on *Inoviridae*, which have lipid membrane-adsorbing coat proteins[276]. The 18 viral populations that were under strong methodological biased were filtered out of the final selection of adsorbing vial populations.

After applying MDA and methodological selection filters, 26 viral populations with predicted adsorbing activity remained (Supplementary Figure 3A). All 26 selected viral populations were highly specific to cell envelopes of a single bacterium, and thus represented phages with a single predicted host. This was despite our allowance of broad host-range at the order level in selecting viral populations, but agrees with a recent report that found that broad host-range phages are rare in gut viromes based on single-cell viral tagging experiments[88].
Notably, 22 of the 26 selected viral populations putatively adsorb to *Proteobacteria* cell envelopes, consistent with recent findings that *Proteobacteria* are the dominant bacterial phylum in global wastewater treatment plant microbiomes and are abundant in wastewater influent[277,278]. It is thus likely that *Proteobacteria* phages are common in wastewater microbiomes, which supports our findings.

Selected viral populations are rare and similar to Proteobacteria phages

Next, we examined the final selection of 26 viral populations with adsorption predictions. Their relative abundance in the virome ranged from 0.0001% to 1%, covering the spectrum from relatively rare to relatively abundant in the virome before and after MDA (Figure 2A). Characterisation of the ORFs by direct homology searches showed that the majority of ORFs (64% of total) in selected viral populations had no significant similarity to protein sequences in the National Institute for Biotechnology Information (NCBI) non-redundant (nr) database[279] (Figure 2B). From these findings we concluded that AdsorpSeq can be used to identify adsorption hosts of both common and rare uncharacterised environmental phages.

To assess AdsorpSeq host predictions, we performed a whole genome clustering of all viral populations (including those without adsorption predictions) with all characterised phages whose genome sequences were available in the NCBI bacterial and viral RefSeq V85 database[153] (Figure 2C). First, we observed that the contigs of the viral populations identified with AdsorpSeq clustered together, confirming our contig binning approach. Second, most of the 22 selected viral populations that were predicted to adsorb to *Proteobacteria* species were similar to characterised *Proteobacteria* phages. This supports the notion that the viral populations that were detected with AdsorpSeq on the cell envelopes of specific bacterial pathogens reflect naturally occurring phages capable of infecting these bacteria.

Selected viral populations with similarity to characterised phages

To further assess the phage-host associations detected by AdsorpSeq, we searched for matches between contigs in selected viral populations and bacterial CRISPR-Cas spacers. We identified 1.4 million spacers predicted from bacteria in the Pathosystems Resource Integration Centre (PATRIC) database[151] and queried them against contigs in the selected viral populations. No full-length identical protospacers were detected on any of the viral contigs. Two contigs from viral populations adsorbing to *Pseudomonas aeruginosa* contained spacer hits with a single mismatch each. These hits originated from CRISPR-Cas arrays encoded on the genomes of *Aeromonas caviae* and *Tolumonas auensis,* which are members of the same taxonomical class as *P. aeruginosa* (*Gammaproteobacteria*). Several examples have been observed of phage genera that adsorb to differently related bacteria, such as Tequatroviruses[280], Gap227likeviruses[281], and Plpelikeviruses[72,282]. Alternatively, cell envelope adsorption does not necessitate successful infection, as many factors play a role in the completion of the

infection cycle[128], including differences in transcriptional programs between hosts[204] or the presence of molecular defence system [283,284]. No other selected viral populations contained contigs with spacer hits with fewer than five mismatches. As five mismatches corresponds to a ~50% false discovery rate at the species level[92], hits with more mismatches were not considered for analysis.

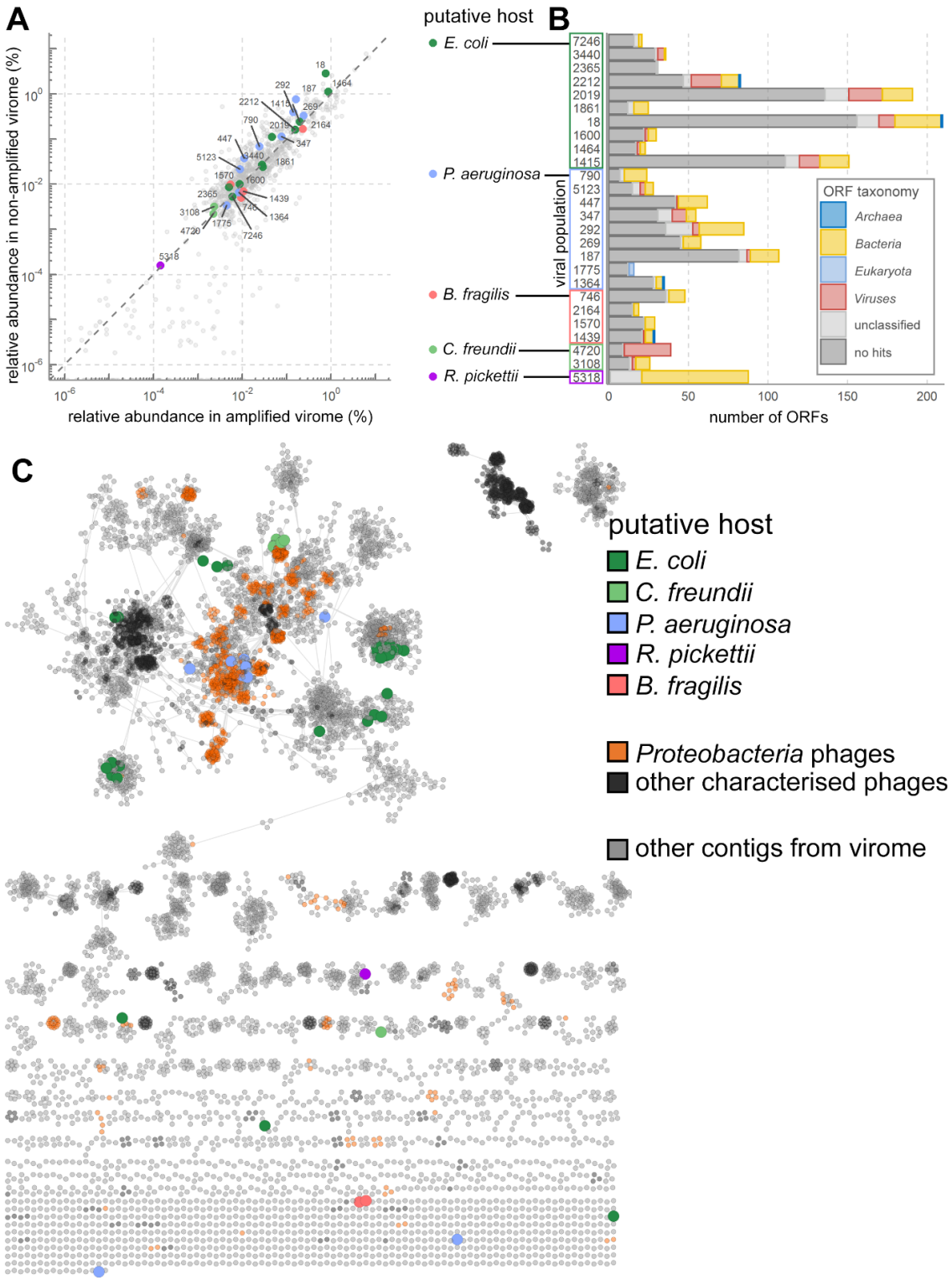


Figure 2: Most selected viral populations represent rare and uncharacterised viral sequences. (A) *Relative abundance of selected viral populations with adsorption predictions in the virome before and after MDA shows that AdsorpSeq is not biased for abundant or rare phage sequences. Numbers next to data points show the viral population number.* ***(B)*** *ORF-level taxonomical predictions using CAT show most ORFs from selected viral populations have no similarities in the NCBI nr protein sequence database (dark grey). Some contigs had database hits but could not be classified because the hits involved proteins from different superkingdoms. These are labelled as unclassified (light grey).* ***(C)*** *The hospital wastewater virome contained a large diversity of uncharacterised phage sequences, as shown by a gene-sharing network of 10,032 viral contigs and all phage genomes in the NCBI viral RefSeq database [153], made using vContact2 [122]. Large coloured contigs represent those in the final selection of 26 putative adsorbing viral populations. Proteobacteria-infecting characterised phages are orange.*

Next, we placed the viral populations identified with AdsorpSeq in the context of characterised phages by using protein sharing networks. We narrowed our search to contigs that contained over five ORFs with known homologs, yielding five selected viral populations with extensive similarity to characterised phages (Figure 3, Supplementary Table 3).

Firstly, *P. aeruginosa*-adsorbing viral population 292 contained a single circular contig on which 15/85 ORFs (17.6%) where similar to proteins from the known *Pseudomonas* phages JBD44, phi297, and YMC11/07/P54_PAE_BP (Figure 3A). Notably, phi297 infects the *P. aeruginosa* strain used in this study[285]. The fifteen shared proteins included the often conserved terminase[123,124] and several adjacent genes (Figure 4A, Supplementary Table 4), stressing the relatedness between these phages. Following earlier practice[286], this protein similarity would classifies these phages into the same taxonomical family. While the conservative CAT ORF-level predictions provided no taxonomic classifications for most ORFs in viral population 292, direct homology searches found significant similarity (BLASTp, e-value $\leq 10^{-5}$) to proteins from *Pseudomonas* bacteria for 45 of the 85 ORFs (53%, Figure 4A, Supplementary Table 4) from viral population 292. This viral population thus represents a novel *Pseudomonas* phage identified in the hospital sewage inlet by AdsorpSeq with *P. aeruginosa* cell envelopes as bait. A second viral population that also adsorbed to *P. aeruginosa* (447) contained a contig of 38,882 bp that had 5/62 ORFs with similarity to *Myxococcus* phage Mx8 (Figure 3B), which translates to less than 5% of 85 Mx8 ORFs. The similarity between viral population 447 and phage Mx8 is thus limited to a small number of genes, which although not adjacent all belong

to the structural section of the Mx8 genome (Supplementary Table 3). We suggest that these limited shared proteins reflect shared gene cassettes[286,287].

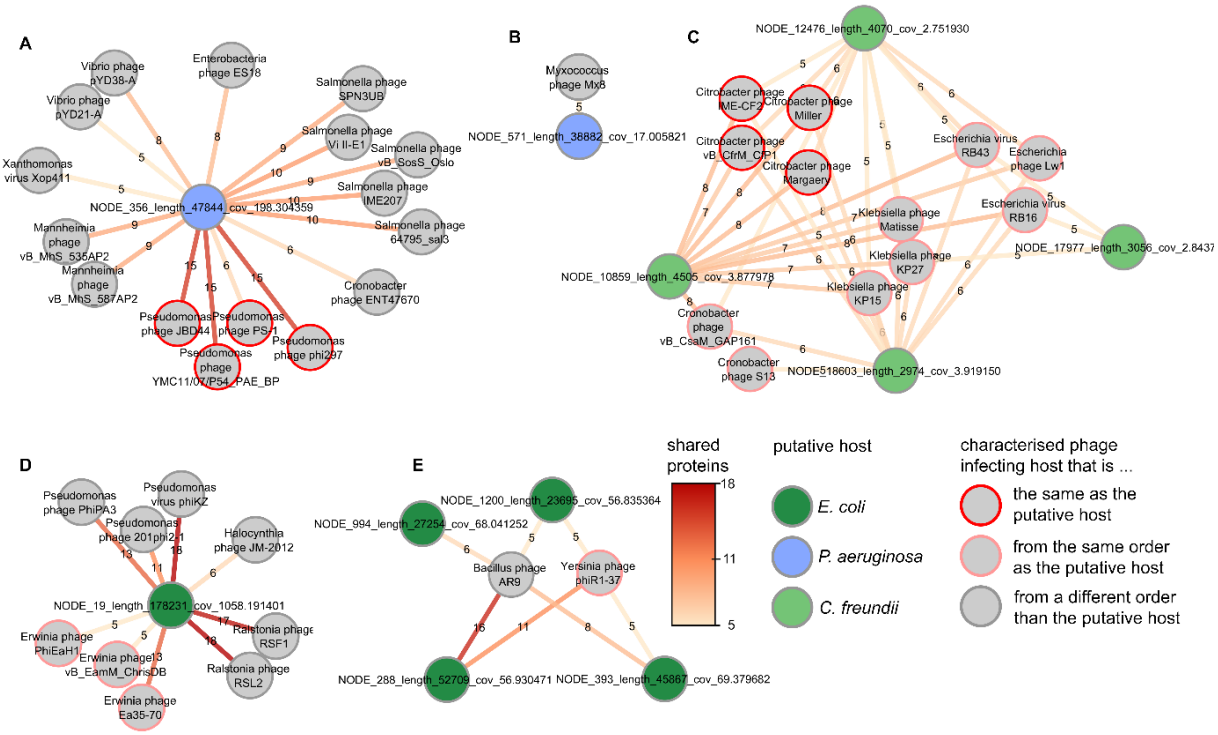


Figure 3: Selected viral populations related to characterised phages.

*Protein-sharing networks of viral populations show their relationships to characterised phages. ORFs from selected viral populations were used in BLASTp searches against proteins of phages in the viral RefSeq database[153]. Bubbles are phages. Edge colour and labels shows similar protein counts (E-value ≤ 10^{-5}). **(A)** P. aeruginosa-adsorbing viral population 292. **(B)** P. aeruginosa-adsorbing viral population 447. **(C)** C. freundii-adsorbing viral population 4720. One additional contig did not share protein similarity to characterised phages. **(D)** E. coli-adsorbing viral population 18. **(E)** E. coli-adsorbing viral population 2019.*

In addition to the *Pseudomonas*-adsorbing viral populations, viral population 4720 represents a *Citrobacter*-adsorbing phage. This viral population contains five short (2974-4505 bp) contigs. The four longest of these had significant protein sequence similarity (BLASTp, e-value ≤ 10^{-5}) to at least five ORFs from several *Citrobacter* phages (Figure 3C), while all five contigs showed full-length sequence similarity to *Citrobacter* phage Margaery (tBLASTx, e-value ≤ 0.001, Figure 4B). Combined, they shared similarity to 25/280 (9%) *Citrobacter* phage Margaery proteins and several other T4-like *Citrobacter* phages (Figure 4B). As these known

Citrobacter phages have genomes over 150,000 bp, we suggest that viral population 4720 may represent fragments of a larger *Citrobacter* phage genome.

Finally, two selected *E. coli*-adsorbing viral populations (18 and 2019) shared five to eighteen ORFs with several known jumbo phages (Figure 3D and E). These included *Yersinia* phage phiR1-37 and *Erwinia* phage Ea35-70, that both infect bacteria from the same taxonomic order as *E. coli* (*Gammaproteobacteria*). Notably, *E. coli* phages like T4[184] and T3[288] need only limited genomic alterations to extend their host range to *Yersinia* species. To gain added insight into the relation of these phages to other jumbo phages, we built a phylogeny of their terminase genes (Figure 4C). Earlier analysis showed that jumbo phages with phylogenetically closely related terminases often infect related hosts[289]. For our analysis, we gathered 148 sequences from the NCBI nr database[279] and 74 jumbo phage proteins from a recent study [8] that had significant homology to viral population 18 and 2019 terminases (BLASTp, e-value $\leq 10^{-5}$). The resulting tree placed viral population 2019 in a branch that mostly held recently described jumbo phages[8], with hosts predicted to belong to the *Firmicutes*. Viral population 18 terminase belonged to a more diverse clade with multiple *Ralstonia* phages, a single *Alteromonadaceae* bacterial sequence (a *γ-proteobacteria* species), and multiple jumbo phages previously predicted to infect proteobacterial hosts[8]. These phages might interact with a surface element that is common to all these hosts. This hypothesis is reinforced by the fact that *Ralstonia* were long assigned to the genus *Pseudomonas* within the γ-proteobacteria, despite differences in membrane composition[290] and lipopolysaccharide structure[291]. Alternatively, in similar fashion to *Enterobacteria* phage phi92[171], they may harbour multiple receptor-adsorbing proteins in their large genomes. Besides the terminase, the circular contig in viral population 18 notably also contained an FtsZ homolog. In *Pseudomonas* jumbo phages, FtsZ-proteins are part of a nucleus-like defence mechanism[200,292], together with a nucleus-forming protein. Interestingly, protein homology searches did not identify similar nucleus-forming proteins in viral population 18, although this population consists of a single circular contig. It may thus contain a yet unknown system like the nucleus-like defence mechanism found in *Pseudomonas* jumbo phages or use the FtsZ homolog in a system that is different altogether.

The results discussed above covered represent 6 of 26 AdsorpSeq-selected viral populations with at least five protein similarities to characterised phages. The ORFs encoded on the remaining 20 selected viral populations represented novel or highly divergent proteins, unrelated to previously characterised phages. Together, our results underscore the ability of AdsorpSeq to uncover environmental phages and their potential host associations, without bias for known or abundant phages.

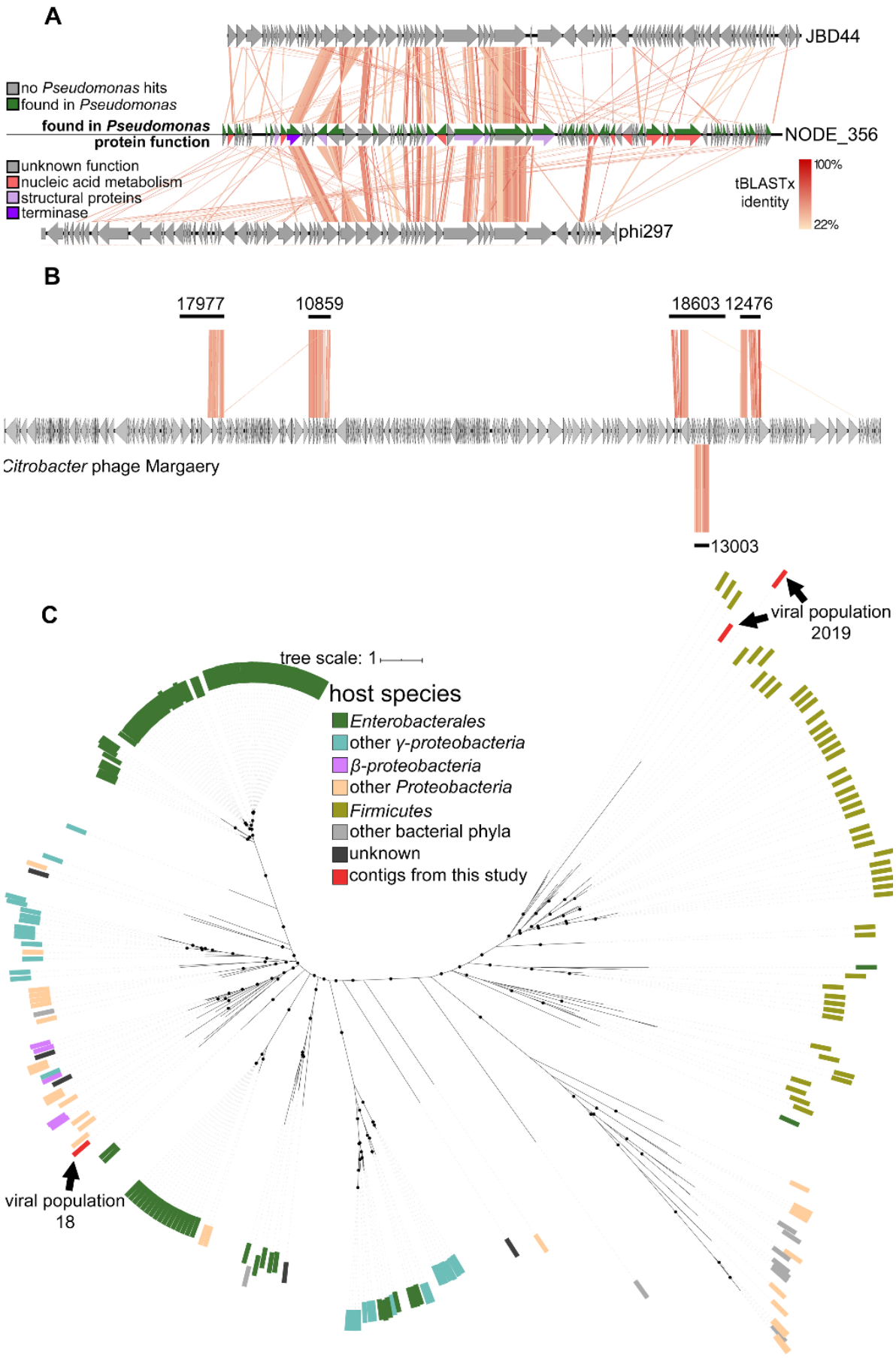

A
JBD44
NODE_356
phi297
no Pseudomonas hits
found in Pseudomonas
found in Pseudomonas protein function
unknown function
nucleic acid metabolism
structural proteins
terminase
tBLASTx identity
100%
22%
B
17977
10859
18603
12476
Citrobacter phage Margaery
13003
C
tree scale: 1
host species
Enterobacterales
other γ-proteobacteria
β-proteobacteria
other Proteobacteria
Firmicutes
other bacterial phyla
unknown
contigs from this study
viral population 2019
viral population 18

Figure 4: Several viral populations and their relations to characterised families. (A) *Similarity of viral population 447, containing contig 356, to Pseudomonas phages JBD44 and phi297. Depicted is a whole genome comparison made using Easyfig[144]. In the line representing contig 356, the top half shows ORFs with BLASTp hit against Pseudomonas bacteria proteins in the NCBI nr database, while the bottom half shows protein function.* ***(B)*** *Similarity between five contigs from C. freundii-adsorbing viral population 4720 and T4-like Citrobacter phage Margaery, as shown by genome comparisons made using Easyfig[144]. Numbers indicate contig numbers, contig 13003 was placed below phage Margaery as it overlaps with contig 18603. Colours indicate tBLASTx hits and use the same legend as (A).* ***(C)*** *The relation of E. coli-adsorbing viral populations 18 and 2019 to jumbo phages displayed in an unrooted approximate maximum likelihood tree of jumbo phage terminases. Dots on branches represent ultrafast bootstrap support of ≥85[149].*

Conclusion

Despite recent advances in viral metagenomics, linking environmental phage sequences to hosts remains problematic[92]. Here, we presented AdsorpSeq, a rapid method for detecting phage adsorption to host cellular envelopes . AdsorpSeq is easily implementable, as besides sequencing it uses only commonly available laboratory methods, such as cell disruption and agarose gel electrophoresis. In the current study, we validated AdsorpSeq on model bacteriophages. This showed that AdsorpSeq can separate phages based on the presence of a specific host receptor in bacterial cell envelopes. Future improvements to AdsorpSeq could exploit this feature, for instance through heterologous expression of certain bacterial features of interest and uncovering phages that adsorb to them. Therewith, AdsorpSeq could aid in uncovering interactions between phages and microbial surface molecules like antimicrobial efflux pumps. AdsorpSeq uncouples analysis of phage adsorption and phage infection, potentially allowing for analysis of phage-bacterium interactions that do not result in infection, for instance due to intracellular defence systems[284].

The experiments laid out in this study identified 26 phage-host interactions in a hospital wastewater virome. Most of these were predictions of phages adsorbing to *Proteobacteria,* which is consistent with recent findings that these bacteria are ubiquitous and abundant in global waste water communities[277,278]. Several of the putative adsorbing viral populations represented rare members of the sampled hospital wastewater virome. While rare here, these viruses may

still be important in another time and place as they become transiently dominant through ecological dynamics[293].

Moreover, expanding our knowledge to include rare members of the virosphere is important to address fundamental questions about the evolution of viral genes and genomes[294,295], to identify candidate viruses with potentially promising genomic properties for phage therapy[238], and for targeted monitoring[296]. This underscores the necessity for novel methods such as AdsorpSeq with which phage-host interactions can be rapidly assessed, even for uncommon phages within a complex environmental mixture.

Methods

General data reporting

All metagenomics tools used default parameters, except where explicitly stated otherwise. All graphs were plotted using the ggplot[297] v3.2.1 package in R. All chemicals were obtained from Sigma-Aldrich, unless explicitly stated otherwise.

Data availability

All genomic data has been uploaded to the European Nucleotide Archive (ENA) under project PRJEB37817. Reads are available under ENA accession numbers ERS4427880- ERS4427890, while the cross-assembled contigs are available under ENA accession number ERZ1305919.

Bacterial cultivation and phage stock preparation

We tested the validity of AdsorpSeq with two model phages and their hosts. These were *Escherichia* phage λ (DSMZ #4499) infecting *Escherichia coli* K12 BW25113 (DSMZ #27469) and *Salmonella* phage P22 (DSM #18523), infecting *Salmonella enterica* subsp. *enterica* serovar Enteritidis S1400/94 (hereafter called *Salmonella enterica*)[298]. For additional tests of phage λ specificity, we used *E. coli* JW3996 (ΔLamB) from the Keio strain collection[269], which lacks the λ protein receptor[266]. For subsequent AdsorpSeq application on a hospital wastewater virome, we used nine bacterial strains. These were *Acinetobacter baumannii* (DSMZ #300007), *Klebsiella pneumoniae* (ATCC #11296), *Ralstonia pickettii* (DSMZ #6297), *Pseudomonas aeruginosa* PA01 (DSMZ #22644), *Fusobacterium necrophorum* D12, *Citrobacter freundii* 4_7_47CFAA, *Escherichia coli* 4_1_47FAA , *Bacteroides fragilis* 3_1_12, and *Bacteroides dorei* 5_1_36/D4. The latter five strains were provided by the reference catalogue of the Human

Microbiome Project (HMP) from the University of Guelph in Guelph, Canada. All bacteria, except for the obligate anaerobic *Bacteroides* and *Fusobacterium* strains, were aerobically cultivated in lysogeny broth (LB) at 37°C while under agitation. The obligate anaerobic strains were cultivated at 37°C in anaerobic Columbia broth. This medium was made anaerobic by boiling and cooling under a stream of nitrogen gas, followed by dispensation in bottles which were closed with rubber stoppers and aluminium crimp caps. Bottles were subjected to three rounds of vacuum and nitrogen gas and autoclaved at 121°C for 20 minutes.

Phage λ and P22 stocks were produced with the soft-agar overlay method as described[299]. To determine phage titres, an aliquot of 0.1 ml exponentially growing bacterial culture was added to 5 ml 0.7% (w/v) LB agarose. This mixture was layered on top of 1.5% (w/v) LB agar and allowed to dry. Phage stock was diluted in a series of 10-fold dilutions using SM buffer (100 mM NaCl, 8 mM $MgSO_4$ x 7 H_2O, 50 mM Tris-HCl pH 7.5) and 10 µl of each dilution was placed on the plates. After 16 h incubation at 37°C, plaques were counted.

Bacterial cell envelope isolations

To isolate bacterial cell envelopes, bacteria were first grown overnight (aerobic) or for 3 days (anaerobic) as described under "bacterial cultivation and phage stock preparation". The resulting cultures were centrifuged at 10,000 x *g*, 4°C for 15 minutes. Supernatant was discarded and cell pellets were washed with the original volume of lysis buffer (50 mM Tris HCl pH 7.5, 2 mM $MgCl_2$). This was centrifuged again, the supernatant was discarded, and the pellet was re-suspended in 4 volumes of lysis buffer per gram wet cell weight with the addition of 1 tablet of cOmplete EDTA-free protease inhibitor. Cells were lysed by thrice passing the suspension through a model CF1 Cell Disruptor (Constant Systems) at 1.5 kBar. After removal of cell debris by centrifugation at 12,000 x *g*, 4°C for 15 minutes, cell envelopes were collected by centrifugation at 225,000 x *g* ,4°C for 1 hour. The supernatant was discarded, cell envelope pellets were re-suspended in lysis buffer and centrifuged again. Soluble proteins were removed by resuspension of the cell envelope pellet in 200 mM NaCl + 20 mM Tris HCl, pH 7.5 and a third centrifugation. Supernatant was removed and cell envelope pellets were dissolved in 20 mM Tris HCl pH 7.5 + 200 mM NaCl at a concentration of 10 mg/ml and stored at -20°C until further use.

Agarose assays with model phages

AdsorpSeq was first tested using phage λ and P22 and bacterial cell envelope suspensions from their hosts. Equal volumes of 10^{11} pfu/ml phage titre and 10 mg/ml cell envelope suspension were mixed. To allow phages to adsorb to bacterial cell envelope suspensions, the mixture was incubated at room temperature for 20 minutes. Bound and unbound phages were separated by applying the mixtures on a 1% (w/v) agarose gel and applying a current of 20 V/cm for 20 minutes using a Mupid One gel electrophoresis system (Eurogentec). The slots at the top of the gel, which contain bound phages, were cut out of the gel using a fresh scalpel knife and DNA was isolated using a Zymoclean gel DNA recovery kit (Zymo Research). The recovered DNA was quantified with a Qubit dsDNA HS assay kit and Qubit fluorometer (Thermo Fisher Scientific).

Hospital virome preparation

Material for the virome consisted of two litres of wastewater influent kindly provided by the Reinier de Graaf hospital wastewater treatment facility in Delft, the Netherlands. To remove debris and cellular material, the material was filtered in stages using coffee filters, 0.45 µm filters, and 0.2 µm filters. Subsequently the virome was concentrated to 70 ml (i.e. approximately 30 times) using a Vivaflow tangential flow filter with a 100 kDa cut-off (Sartorius). The sample was stored at 4°C until further use.

To isolate phage DNA from the virome, viral capsids were first broken by incubating a 1 ml virome aliquot with 10 µg/ml proteinase K and 0.2 % SDS at 56°C for 1 hour. Afterward, DNA was purified by phase separations using consecutively phenol, phenol/chloroform, and chloroform. The aqueous phase was collected and 0.1 volume 3 M sodium acetate (pH 5.2) and 2.5 volumes ice-cold absolute ethanol were added. The mixture was incubated overnight at -20°C and DNA was pelleted by centrifugation at 21,000 x *g* for 15 minutes at 4°C. The pellet was washed with one volume of 70% ethanol, and re-pelleted by repeating the centrifugation. The DNA pellet was dissolved in TE buffer (1 mM EDTA, 10 mM Tris HCl pH 8) and the DNA concentration was ascertained using a Qubit dsDNA HS assay kit and Qubit fluorometer (Thermo Fisher Scientific).

Sample preparation and sequencing

Two sequencing experiments were performed. The first was a confirmation that AdsorpSeq can be used with a mixture of phage λ and P22, while the second applied AdsorpSeq to the hospital

wastewater virome and nine bacterial strains (see "bacterial cultivation and phage stock preparation"). For the first sequencing experiment, phage λ and P22 were mixed at a titre of approximately $1 \cdot 10^8$ pfu/ml each. This mixture was added to cell envelope suspensions of either *E. coli* BW25113 or *S. enterica*, and DNA of bound phages was obtained as described under "agarose assays with model phages". For the second sequencing experiment, the same methodology was used on samples prepared from the hospital wastewater virome and cell envelope preparations. To increase DNA quantities, the samples were subjected to multiple displacement amplification (MDA) using an Illustra genomiphi v3 φ29 polymerase kit (GE Lifesciences) according to the instructions. Primers were removed from the amplified samples using AMpure XP beads (Beckman Coulter) and the final amplicons were eluted in 50 µl of TE buffer. DNA concentrations were determined using a Qubit dsDNA HS assay kit and Qubit fluorometer (Thermo Fisher Scientific). All samples with cell envelope suspensions added to them were prepared in biological quadruplicates. The whole procedure was repeated for the samples including both cell envelopes and viromes, for a total of two technical duplicates that each consisted of biological quadruplicates. Virome-only controls only had biological duplicates. For both experiments, these virome-only controls consisted of DNA isolated from the phages both before and after MDA.

Library preparation and sequencing of all samples was performed at the Utrecht sequencing facility (USEQ) in Utrecht, the Netherlands. Libraries were prepared using the TruSeq DNA nano kit (Illumina), and samples were sequenced on a NextSeq500 run with 2x150 bp paired reads (Illumina).

Sequencing data analysis

To increase the read quality from our sequencing experiments, Illumina reads were quality trimmed, poly-G tails were removed, and remaining adapters were removed using fastp v0.20.0[300] (options -g, -x). For the application of AdsorpSeq on the hospital waste water virome, trimmed reads of the unamplified virome and all bacterial samples were cross-assembled into 1,013,501 contigs using metaSPAdes v3.11.0[301]. Reads from all samples were mapped to the genomes of the model phages and their hosts (model phage experiment) or the cross-assembled contigs (hospital waste water virome experiment) using the Burrows-Wheeler Aligner v0.7.12-r1039[302]. The number of reads that mapped to each genome or contig were determined using the idxstats tool in samtools v1.454[303].

For the experiment using hospital waste water, mean read depths of each contig in each sample were determined using the JGI summarize bam contig depths tool in metaBAT v2.12.1[304]. Additionally, all reads from the amplified virome were mapped against the contigs to determine which contigs were selected for by the amplification procedure. Contigs belonging to the host genomes were identified by a megaBLAST search against the host genomes using BLAST v2.6.0+[145]. The 12,496 contigs with over 50% coverage and over 50% identity were removed as host contigs. Subsequently, contigs were taxonomically annotated using the contig annotation tool (CAT) v5.0.3[305], which uses homology searches of ORFs against the National Centre for Biotechnology Information (NCBI) non-redundant protein database (nr)[279] to predict contig taxonomy. CAT used prodigal v2.6.3[135] to predict ORFs. Contigs with a superkingdom classification of "Viruses" and a score of 1 (which indicates that all predicted ORFs in a contig were classified in that superkingdom), those that had other superkingdom classifications with scores below 1, and those that could not be classified at all were selected for further analysis. This resulted in a dataset of 13,032 (putative) viral contigs. Putatively completed circular contigs were identified using a custom script that checked whether the start and end of a contig contained identical sequences, following earlier studies[120,306].

We binned the selected contigs according to their tetranucleotide usage patterns and their read depth patterns across the nine bacterial samples using metaBAT v2.12.1[304]. Because viral genomes are smaller than the bacterial genomes for which metaBat was originally built, we lowered the minimal contig length allowance to 2,500 bp and the minimum bin size to 10,000 bp. In addition, we decreased the minimum mean coverage necessary in each sample for binning to 0.001 to allow for contigs that might not be present in all samples (i.e. options -m 2500, -s 10000, and -x 0.001). Contigs that metaBAT could not bin, whether due to low coverage or short length, were discarded. We binned genomic fragments into viral populations[307] based on similarity in tetranucleotide usage and abundance patterns by using metaBat resulting in 1158 viral populations containing a total of 6,572 contigs.

Selection of overrepresented viral populations

To select putative cell envelope-adsorbing viral populations, the abundance per viral population in each sample was calculated with the following formula:

$$\text{abundance}_{\text{viral population}} = \frac{\sum\left(\text{read depth}_{\text{viral population}} \cdot \text{contig length}_{\text{viral population}}\right)}{\sum \text{contig length}_{\text{viral population}}} \quad \text{(Equation 1)}$$

This abundance was then divided by the total abundance of all viral populations in the sample and expressed in a percentage to obtain the relative abundance for each viral population in each sample. For each viral population, the highest relative abundance across all cell envelope-treated samples was divided by the next highest relative abundance from a sample that used cell envelope suspensions from different taxonomical order. Positive outliers among the fractions between the top two samples of all viral populations were determined with the following formula:

$$\text{outlier} \geq 75\%_{\text{all viral populations}} + \text{interquartile range}_{\text{all viral populations}} \cdot 1.5 \qquad \text{(Equation 2)}$$

The boundary for inclusion as putative adsorbing viral population established by this was 1.58 (Supplementary Figure 4A). A total of 123 putative adsorbing viral populations that constituted as outliers were selected.

MDA and methodological selection filters

To determine the extent that viral populations were selected for by MDA, we divided the relative abundance of each viral population in the post-MDA virome by that in the pre-MDA virome. To ascertain the relationship between phage taxonomy and MDA selection factor, we plotted the values of all viral populations with a CAT classification at the family level that belonged to a phage lineage. From the resulting MDA selection factors, positive outliers were determined using equation 2. The 79 viral populations with MDA selection factors above 3.8, as defined by equation 2, were discounted (Supplementary Figure 4B).

In addition to the MDA selection factor, we also determined which viral populations were universally selected for by AdsorpSeq. For this, we calculated a methodological selection factor by dividing the relative abundance of each viral population per sample by the relative abundance in the amplified virome. As a result, each viral population had one methodological selection factor for each cell envelope-treated sample for a total of nine per viral population. After determining positive outliers with equation 2, the 18 viral populations which were above the outlier threshold of 3.43 in all nine samples were removed from the dataset (Supplementary Figure 4C). After accounting for these two biases, our dataset contained 26 putatively adsorbing viral populations which combined contained 83 contigs.

Analysis of putatively adsorbing viral populations

To uncover the extent of viral 'dark matter' among the putatively adsorbing viral populations, we plotted the ORF-specific taxonomical classifications obtained with CAT[305] of all ORFs

found in putative adsorbing viral populations. While CAT is built to classify contigs, it first predicts ORF-specific taxonomy and subsequently uses these results to predict contig taxonomy. Relatedness of the putative adsorbing viral populations and characterised phages was established by performing a BLASTp all vs all on a dataset composed of all ORFs from putative adsorbing viral populations and all ORFs from characterised phages in the NCBI bacterial and viral RefSeq V85 database[153]. BLASTp searches were performed using diamond v0.9.29.130[137], with a bit score significance cut-off of 50. Links between viral populations and characterised phages were displayed using Cytoscape v3.7.2[308]. For further analysis of the links between putative adsorbing viral populations and known phages, we performed a taxonomical assignment on all selected contigs using vContact2 v.0.9.8[122], while the resulting network was visualised using Cytoscape v3.7.2[308].

Contigs from putative adsorbing viral populations were annotated using PROKKA v1.11 [143], with the metagenomic option enabled and with both the viral and bacterial settings (i.e. options--metagenome and both --kingdom Bacteria and --kingdom Viruses). In addition, distant homologs to ORFs were identified by performing searches against prokaryotic viral orthologous groups (pVOGs)[309] using hmmsearch v3.1b2[136]. To compare contig sequences to characterised phage genomes they were aligned using Easyfig v2.2.3[144], which used BLAST v2.9.0+[145] to perform tBLASTx searches.

CRISPR-Cas spacer analysis

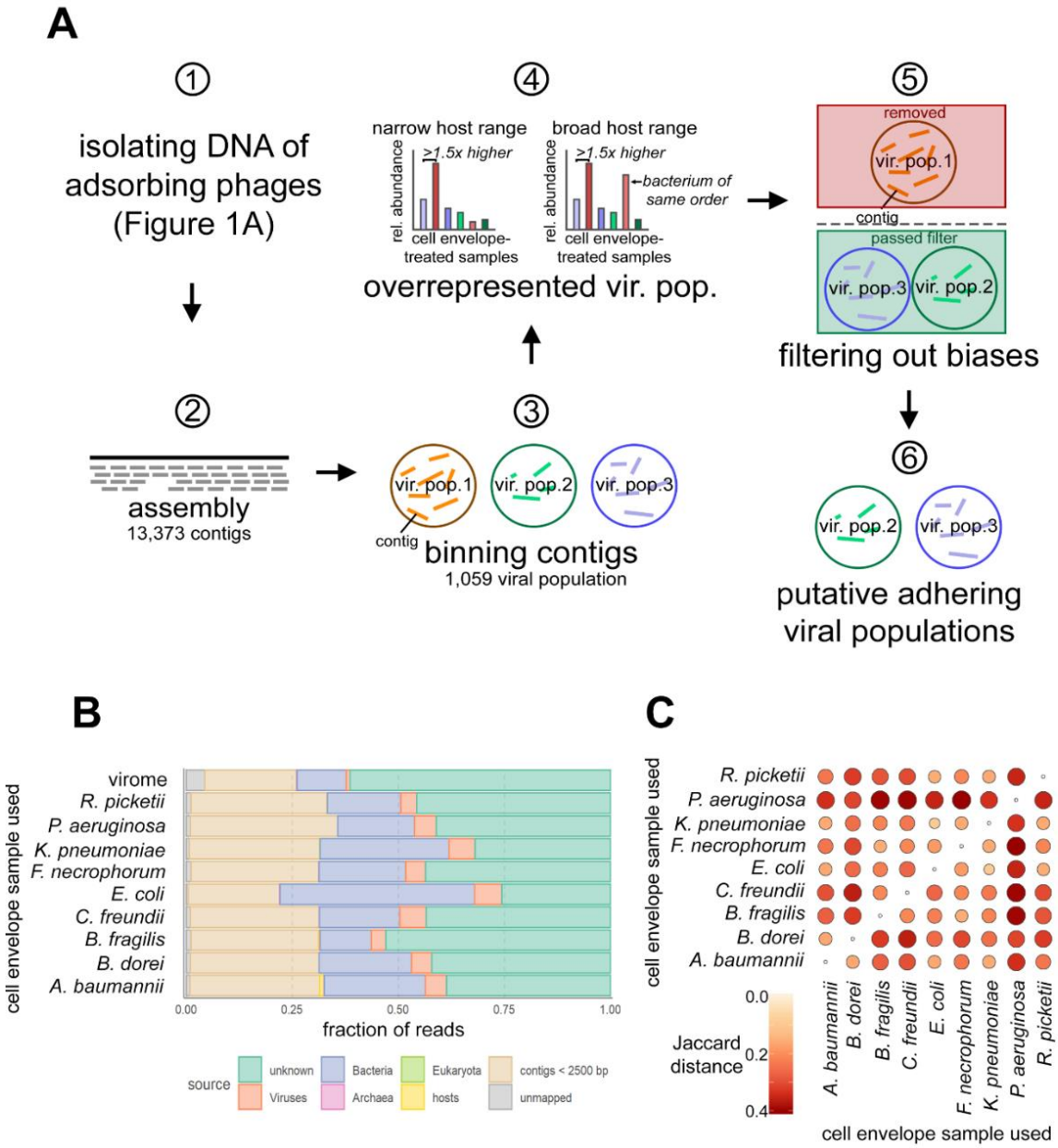
To search for CRISPR-Cas spacer hits against the putative adsorbing viral populations, spacers were identified in all bacteria Pathosystems Resource Integration Centre (PATRIC) database[151] (accessed February 2019) using detect 2.2.1. This resulted in a dataset of 1,473,418 spacers. These spacers were used as query against contigs from putative adsorbing viral populations in a BLASTn with the option for short sequences enabled (option -task blastn-short) in BLAST v2.9.0+[145]. Only spacer hits with fewer than five mismatches were taken into consideration. In practice, there were no spacer hits with between one and five mismatches.

Jumbo phage terminase phylogeny

Some of the putative adsorbing viral populations contained (fragments of) jumbo phage genomes. It was previously shown that in a terminase phylogenetic tree these phages largely clustered according to host[289]. Three terminases that we identified in the putative adsorbing viral populations were used as query for a BLASTp search against the NCBI nr-database using

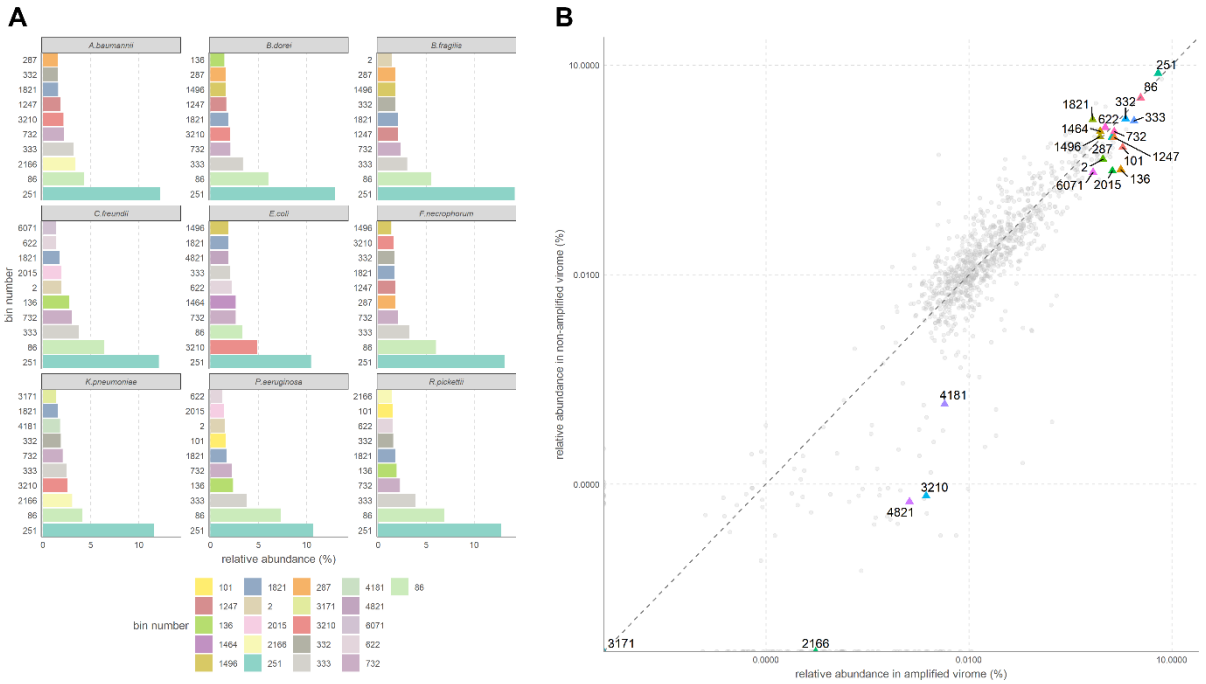
the NCBI webserver[154] on January 14 2020. In addition, the three terminases were used as query for a BLASTp search against jumbo phage proteins from Al-Shayeb *et al*[8]. All 222 total hits above a bit score threshold of 50 were combined with the three queries. Duplicated sequences were removed and the remainder was aligned using Clustal Omega v1.2.1[141]. Aligned positions that consisted of more than 90% gaps were trimmed with trimAl v1.2[146] (option -gt 0.1). A maximum likelihood tree was constructed using IQ-Tree v1.6.12[132] using model finder[148] and performing 1000 iterations of both SH-like approximate likelihood ratio test and the ultrafast bootstrap approximation (UFBoot)[149]. In addition, ten iterations of the tree were separately constructed, as has been recommended[147] (IQ-Tree options -bb 1000, -alrt 1000, and --runs 10). The iteration with the best log-likelihood score, for which model finder selected LG+F+R6, was used for analysis. The tree was visualised using interactive Tree of Life v5.5[150]. In addition, we selected the putative FtsZ homolog from viral population 18 (ORF 17) and used it to query the NCBI nr-database using the NCBI webserver[154] on January 14 2020. This did not result in significant hits.

Supplementary Figures

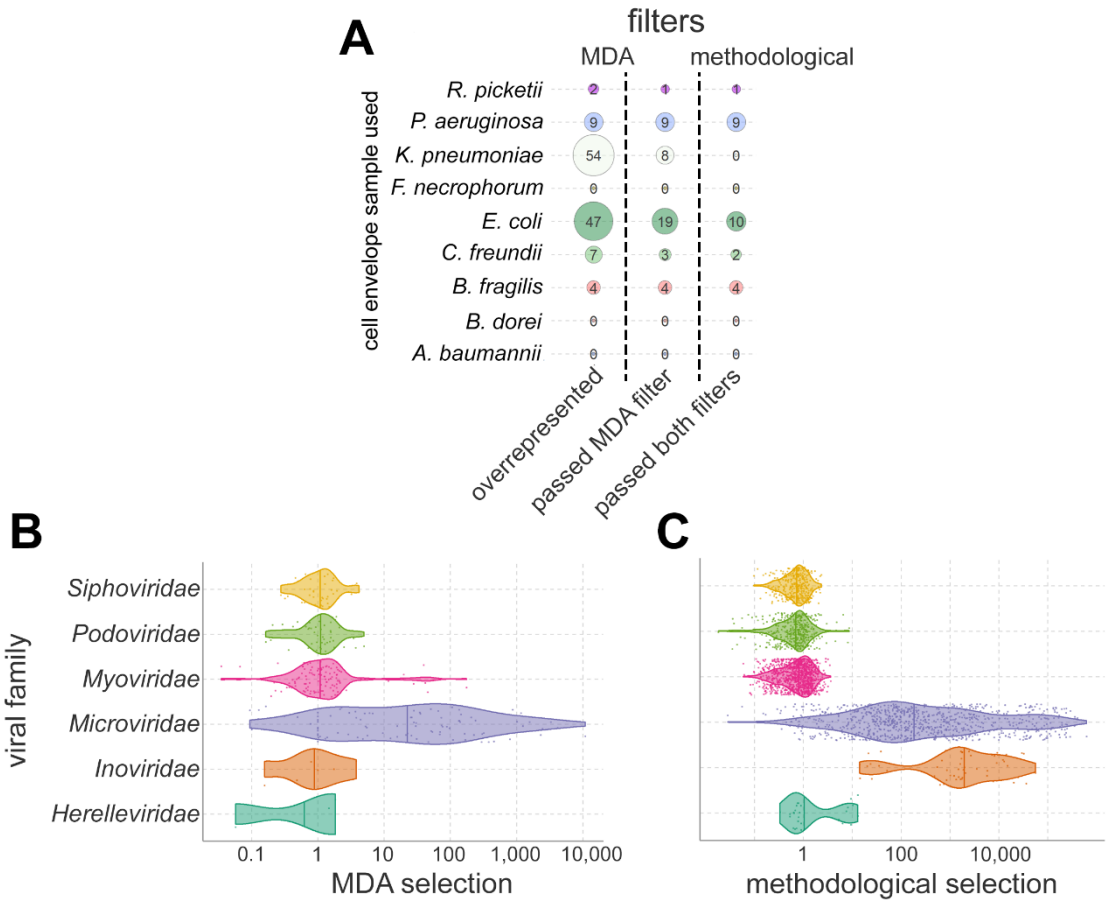


Supplementary Figure 2: Application of AdsorpSeq to a hospital wastewater virome (A) *Schematic of the approach taken to apply AdsorpSeq to an environmental sample. Step (1) consists of the process depicted in Figure 1A. Subsequently, data analysis consisted of (2) assembly of read-level data into contigs, (3) binning contigs into viral populations based on their relative abundance across the samples, (4) selection of viral populations that were overrepresented in one sample, (5) remove viral populations under multiple displacement amplification (MDA) or methodological bias, and (6) the final selection of viral populations with putative adhering capabilities.* ***(B)*** *For contigs over 2500 bp, a majority of reads across the all but one of the samples map to unknown or viral contigs. Stacked bar charts of the*

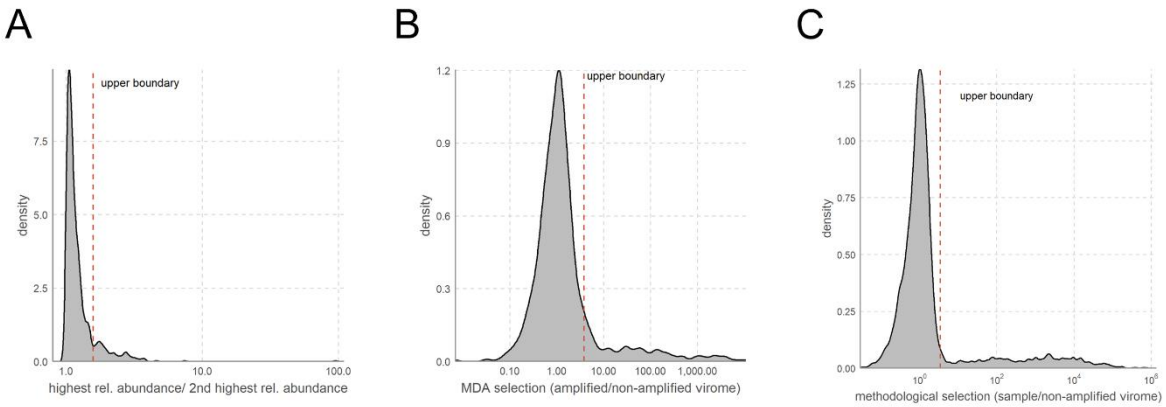
fragment of reads from each sample that are unmapped, map to short contigs below 2,500 bp, map to one of the nine bacterial hosts, map to contigs that are classifiable at the superkingdom level, and map to unknown/unclassifiable contigs. ***(C)*** *Selection of phages by AdsorpSeq is dependent on the species of bacterial cell envelope that is used. A correlogram of the Jaccard distance between the membrane-treated samples, based on 1,058 viral populations. Higher Jaccard distance means greater dissimilarity. Size and colour indicated Jaccard dissimilarity.*



Supplementary Figure 3: The most abundant genome viral populations across the cell envelope-treated samples are similar. ***(A)*** *Bar charts showing the ten most abundant genome viral populations in each of the nine cell envelope-treated samples and their relative abundances.* ***(B)*** *The relative abundances of the genome viral populations depicted in (A) in the virome before (y-axis) and after (x-axis) multiple displacement amplification. Numerical labels refer to the genome viral populations.*



Supplementary Figure 3: Selection of putative adhering viral populations from a hospital wastewater virome. ***(A)*** *After filtering for MDA and methodological selection, 26 viral populations with putative adsorbing capacity were selected from the virome. Bubble area shows number of viral populations after sequentially selecting based on being overrepresented in one sample, filtering for MDA selection, and filtering for methodological selection. The third column shows the selected viral populations with putative adhesion activity.* ***(B)*** *Microviridae are highly selected for by MDA, as shown by viral populations with positive CAT predictions within a bacteriophage family. MDA selection is defined as the ratio in relative abundance between the virome before and after MDA. Points in the charts represent individual viral populations.* ***(C)*** *Inoviridae and Microviridae undergo high methodological selection, which is defined by the ratio between the relative abundance of a viral population in each sample and in the amplified virome. As there were nine cell envelope-treated samples, each viral population is represented in the plot by nine data points.*



Supplementary Figure 4: Distributions used in selection of viral populations that represent adhering viral populations. (A) *Distribution of the ratios between the cell envelope-treated sample with the highest and second highest relative abundance among the viral populations. All viral populations with a ratio above the red dashed line labelled "upper boundary" were selected as being overrepresented in a sample.* ***(B)*** *Distribution of the ratios in relative abundance between the virome before and after multiple displacement amplification (MDA). All viral populations with a ratio above the red dashed line labelled "upper boundary" were under strong MDA selection.* ***(C)*** *Distribution of the ratios between the relative abundances in the samples and in the non-amplified virome for each viral population. Each viral population is represented with nine datapoint in this plot; one for each cell envelope-treated sample. All viral populations with a ratio above the red dashed line labelled "upper boundary" were under strong methodological selection.*

5

Development of styrene maleic acid lipid particles (SMALPs) as a tool for studies of phage-host interactions

"You certainly usually find something, if you look,
but it is not always quite the something you were after."
The Hobbit; J.R.R. Tolkien

Submitted as: de Jonge, P.A.; Smit Sibinga, D.J.C.; Boright, O.A.; Costa, A.R.; Nobrega, F.L..; Brouns, S.J.J.; Dutilh, B.E.; *Development of styrene maleic acid lipid particles (SMALPs) as a tool for studies of phage-host interactions*

Abstract

The infection of a bacterium by a phage starts with attachment to a receptor molecule on the host cell surface by the phage. As receptor-phage interactions are crucial to successful infections, they are major determinants of phage host-range and by extension of the broader effects that phages have on bacterial communities. Many receptor molecules, particularly membrane proteins, are difficult to isolate because their stability is supported by their native membrane environments. SMALPs, a recent advance in membrane protein studies, are the result of membrane solubilizations by styrene maleic acid (SMA) co-polymer chains. SMALPs thereby allow for isolation of membrane proteins while maintaining their native environment. Here, we present the development of styrene maleic acid lipid particles (SMALPs) as a tool to isolate and study phage-receptor interactions. In this study we show that SMALPs produced from taxonomically distant bacterial membranes allow for receptor-specific inactivation of, and binding to, several model phages that span the three largest phage families. After characterizing the effects of incubation time and SMALP concentration on the activity of three distinct phages, we present evidence that the interaction between phages and SMALPs is specific to bacterial species and the phage receptor molecule. These interactions additionally lead to DNA ejection by nearly all particles at high phage titers. We conclude that SMALPs are a potentially highly useful tool for phage host-interaction studies.

Introduction

The first stage of infection by a bacteriophage consists of the viral particle binding to a cognate receptor on the host cell surface[12]. Each phage attaches to a specific receptor molecule, and as a result largely determine the specificity to hosts that is common among characterized phages[128]. Indeed, specificity of phages to their receptors is a useful tool for characterizing bacterial strains[233]. In addition to phage applications, the molecules involved in phage-receptor interactions are often evolutionary hotspots due to their crucial role in successful infections[177,178]. As such, understanding phage-receptor interactions is important for understanding the evolutionary dynamics that govern networks of phages and hosts[310] and the role that phages play in regulating bacterial communities[36,311,312]. Indeed, phage-receptor binding is a widely investigated research theme, and studies of phage-receptor interactions have among others ranged from the mechanics of phage binding and DNA ejection[313–315] to the structural modelling of bound receptor molecules[103,316].

Known phage receptors include a wide variety of diverse molecules on the bacterial cell surface[174,317]. While some phages bind to pili, flagella, or cell capsules, many phage receptors are associated with the bacterial cell envelope[317]. Precisely which molecules serve as receptor molecules largely depends on the diversity of bacterial cell wall structures[173]. Known receptors in Gram-positive bacteria are mostly in the peptidoglycan layer[173,174], while those in Gram-negative bacteria are about equally divided between sugar moieties in lipopolysaccharide (LPS) chains and membrane-incorporated porin and transport proteins[174,317]. Due to their association with the cell membrane, phage receptors, especially membrane proteins, can be challenging to study. Proteinaceous phage receptors are generally incorporated into the bacterial cell membrane and are thereby dependent on this membrane to maintain correct folding[318], which makes them notoriously difficult to isolate[319,320]. This limits the number of phages of which the interaction with a receptor can be studied in detail.

Among novel solutions to facilitate membrane protein studies is styrene maleic acid (SMA)[321]. SMA is an amphipathic co-polymer composed of chains with alternating hydrophobic styrene and hydrophilic maleic acid groups[322]. Upon addition of SMA to lipid bilayers, it incorporates itself into the membrane, which leads to the solubilization of membranes into styrene maleic acid lipid particles (SMALPs, Figure 1a)[321]. In SMALPs, the SMA polymer is wound around a disk of membrane about 10 nm in diameter, with hydrophobic styrene groups intercalated between lipid tail acyl groups and hydrophilic maleic acid groups pointed outward[322]. Membrane proteins may be captured within the confines of SMALP discs, and SMALPs thus allow isolation of bacterial membrane proteins while maintaining their natural lipid environment. Since their initial development for membrane protein studies, SMALPs have been used to obtain three dimensional structures of membrane proteins[323,324], and obtain their biophysical characteristics[321,325]. As SMALPs are useful agents for bacterial membrane isolation, they are also potentially useful for phage-receptor studies.

In this study, we report on the development of SMALPs as a platform to study phage-host interactions. We test whether SMALPs can inhibit phages that infect taxonomically diverse bacteria by binding to diverse membrane-associated receptor molecules. We then show that phages interacting with SMALPs maintain specificity to their receptors, and that they eject their DNA after binding to SMALPs.

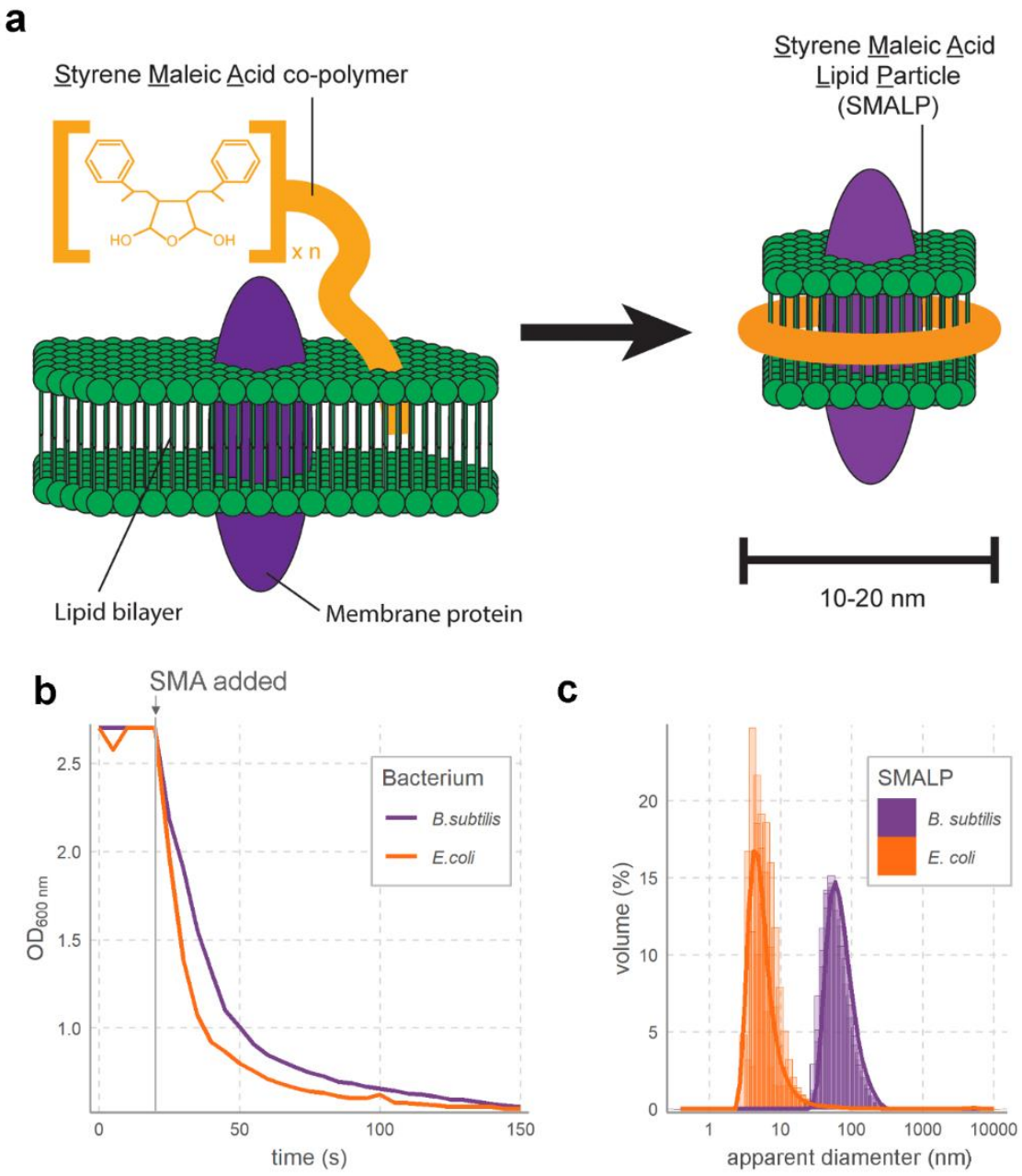


Figure 4: Styrene maleic acid lipid particles (SMALPs) can be made from distinct microbial membranes. ***(A)*** *A schematic of the process by which styrene maleic acid co-polymer incorporates itself into lipid membrane to isolate membrane proteins in SMALPS.* ***(B)*** *SMALPs can be made from Gram-positive and Gram-negative bacterial membranes at roughly equal efficiency. Optical densities of E. coli K12 and B. subtilis 110NA membrane suspensions (10 mg/ml) rapidly decreased upon addition of 20 mg/ml SMA (at grey vertical line) due to dissolution of large membrane particles into SMALPs.* ***(C)*** *SMALPs of B. subtilis are larger than SMALPs of E. coli, as shown by diameter distributions of SMALPs from the two species obtained through dynamic light scattering. Each sample was measured in triplicate, with the replicates plotted on top of each other. Solid lines are the average of the three replicates.*

Results & Discussion

SMALP formation in two taxonomically distant bacteria.

While SMALPs have been developed as a platform to study membrane proteins for over a decade[326], this has mostly focused on *Escherichia coli* membranes. It is consequently unknown whether differences in membrane composition and cell wall structure between bacterial lineages[327] will lead to differences in membrane solubilization by SMALPs. To test this, we started by solubilizing membranes of two taxonomically distant bacteria, one Gram-positive and one Gram-negative, with styrene maleic acid co-polymer. As sample bacteria, we selected *E. coli* and *Bacillus subtilis*. Membrane solubilization of these bacteria, evidenced by diminishing absorbance at 600 nm upon SMA addition, showed similar rates of SMALP formation in the two bacterial membranes (Figure 1b). Dynamic light scattering (DLS) revealed that the Gram-negative *E. coli* membranes formed uniform particles of roughly 10 nm in diameter, as is normally observed for SMALPs (Figure 1c)[321]. Interestingly, DLS of Gram-positive *B. subtilis* SMALPs showed larger particles of 50-100 nm (Figure 1c). Previous studies reported that SMALP size is dependent on the ratio between polar and apolar groups in the SMA polymer[328] and the ratio between polymer and membrane[329,330]. While no extant studies report on the effects of membrane composition from different bacteria on SMALP size, distinct lipid composition of *E. coli* and *B. subtilis* membranes[331] may lead to differences in SMALP sizes as we observed. Beyond lipid composition, the 40 nm thick outer peptidoglycan layer[332] that is attached to the outer membrane of *B. subtilis* by lipoteichoic acids makes cell envelopes much thicker than those of *E. coli*[333]. Since DLS measurements assume a spherical object, this could further contribute to the larger apparent diameter of *B. subtilis* SMALP.

SMALPs inactivate lytic phages.

Next, we tested whether SMALPs from both Gram-positive *B. subtilis* and Gram-negative *E. coli* bacteria inhibit the lytic activity of three model phages (Table 1). As model phages we selected one member from each of the three major families of tailed phages (*Myo*-, *Podo*- and *Siphoviridae*), as the distinct morphologies of these families differentiates their binding mechanisms[12]. The three model phages, *Podoviridae Bacillus* phage φ29, *Siphoviridae Escherichia* phage λ and *Myoviridae Escherichia* phage T4, each bind to distinct receptor molecules. *Bacillus* phage φ29, like many phages infecting Gram-positive bacteria[173], binds to cell wall teichoic acids. *Escherichia* phage λ exclusively binds to a maltose porin protein LamB,

while *Escherichia* phage T4 binds to both outer membrane protein C (OmpC) and glucose moieties in lipopolysaccharide (LPS) chains. While some phages bind to other structures than our model phages, most notably pili and flagella[12], the use of SMALPs to study phage-host interactions will evidently have to focus on membrane-associated structures. Our model phages additionally allowed us to test whether beside proteins, LPS and peptidoglycan layers attached to SMALPs are available to phages.

Table 1: Characteristics of model phages used to test the inhibitory effect of SMALPs.

Phage	Family	Host used	Classification	Receptor(s)
λ	*Siphoviridae*	*E. coli BW25113*	Gram-negative	LamB protein [268]
T4	*Myoviridae*	*E. coli BW25113*	Gram-negative	OmpC protein/LPS [334]
φ29	*Podoviridae*	*B. subtilis 110NA*	Gram-positive	Teichoic acids [335]

As preliminary examination of phage binding by SMALPs, we tested phage lytic capability in liquid cultures after treatment with SMALPs prepared from 10 mg/ml bacterial membrane. Addition of SMALPs to liquid bacterial cultures infected with phage at titers of 10^4 to 10^6 pfu/ml phage (starting multiplicity of infection: 10^{-4} to 10^{-2}, Figure 2a) resulted in complete inhibition of phage lytic activity. Differences in the titer at which we observed complete inhibition of phage lysis suggested varying sensitivities to SMALPs among the three phages. Because these phages differ greatly in infection dynamics (e.g. burst sizes from 100-1000)[335–337], and because liquid assays are nonquantitative[338], we next quantified SMALP-phage interactions through plaque counts after SMALP treatment. Viable phage counts at different time intervals revealed that 1 to 0.1% of 10^6 pfu/ml solutions remained after five minutes of incubation with SMALPs (Figure 2b). Longer incubations showed a slower rate of decrease in viable phage populations, especially for T4 and φ29, while after 20 minutes decreases in viable phage counts largely plateaued. SMALP-phage interactions thus seem to proceed in a two stage process, with a rapid initial phase followed by a slowed secondary phase, similar to the two-stage adsorption dynamics that are characteristic of phages binding to live host cells[90].

In addition to time-dependent dynamics, we measured the effect of SMALP concentrations on phage inhibition (Figure 2c). Undiluted SMALP stocks used in these experiments had been prepared from 10 mg/ml membrane suspensions and 20 mg/ml SMA, but as some of these materials are removed in the SMALP production process, these values do not accurately represent SMALP concentrations. Hence, in the following experiments dilution factor was used instead of concentration counts. Experiments with increasingly diluted

SMALPs confirmed that a positive relationship existed between SMALP concentration and phage inhibition. Of the three model phages, φ29 was most sensitive to SMALPs, as 10,000 times diluted SMALP stocks still inhibited about 90% of a 10^6 pfu/ml phage φ29 solution. As φ29 binds to teichoic acids on the *B. subtilis* cell surface, its sensitivity to SMALPs indicates that the lipoteichoic acid linkages between membrane and peptidoglycan layers[339] are maintained in the *B. subtilis* SMALP solutions. Phages λ and T4 exhibited lower sensitivity to SMALPs than φ29, with SMALP dilutions of a 1,000-fold or more having little effect on phage populations. While at high SMALP dilutions the behavior of these phages is similar, at low dilutions T4 is about two orders of magnitude less sensitive than λ. This likely reflects their disparate binding dynamics, as λ engages in a single interaction with a LamB protein[340] whereas T4 requires up to four independent interactions of tail fibers to receptors for successful infection[341].

If each T4 binding occurs with a separate SMALP particle, likely to be the case due to small SMALP diameters, T4 would thus need fourfold more SMALPs to be inactivated. The same quantity of SMALPs will therefore bind more λ than T4 particles, leading to higher inhibition for λ. Meanwhile, the fact that φ29 exhibits the highest sensitivity to SMALPs implies that teichoic acids are present in SMALPs. Teichoic acids are present in larger numbers in the cell wall than protein receptors such as LamB and OmpC, to which λ and T4 bind. This higher abundance of φ29 receptors increases the number of binding events for φ29 and thus phage inactivation. Together, these results show that for a wide variety of phages that infect taxonomically distant hosts, SMALPs are a potential platform for the study of phage-receptor.

SMALP-mediated phage inactivation is receptor-specific.

After establishing that phages can be inhibited by binding to SMALPs, we focused on the specificity of the SMALP-phage interaction. Preliminary specificity determinations consisted of phage spot tests after treatment with SMALPs made from various bacterial membranes (Figure S1a). Although phage λ at titers of 10^6 pfu/ml still formed plaques after treatment with *E. coli* K-12 SMALPs, the amount of plaques was lower than for non-treated phage. As an indication of specificity, treatment with SMALPs made from an *E. coli* mutant without the LamB receptor (ΔLamB) did not result in an observable decrease in plaque formation. Spot tests of phage T4 at titers of 10^5 pfu/ml with *E. coli* K-12 SMALPs resulted in a near absence of plaques, although a tenfold lower phage titer was needed to produce the effect than for phage

λ. This is in line with the results from the SMALP concentration experiments that indicated phage λ to be more sensitive to SMALPs than T4.

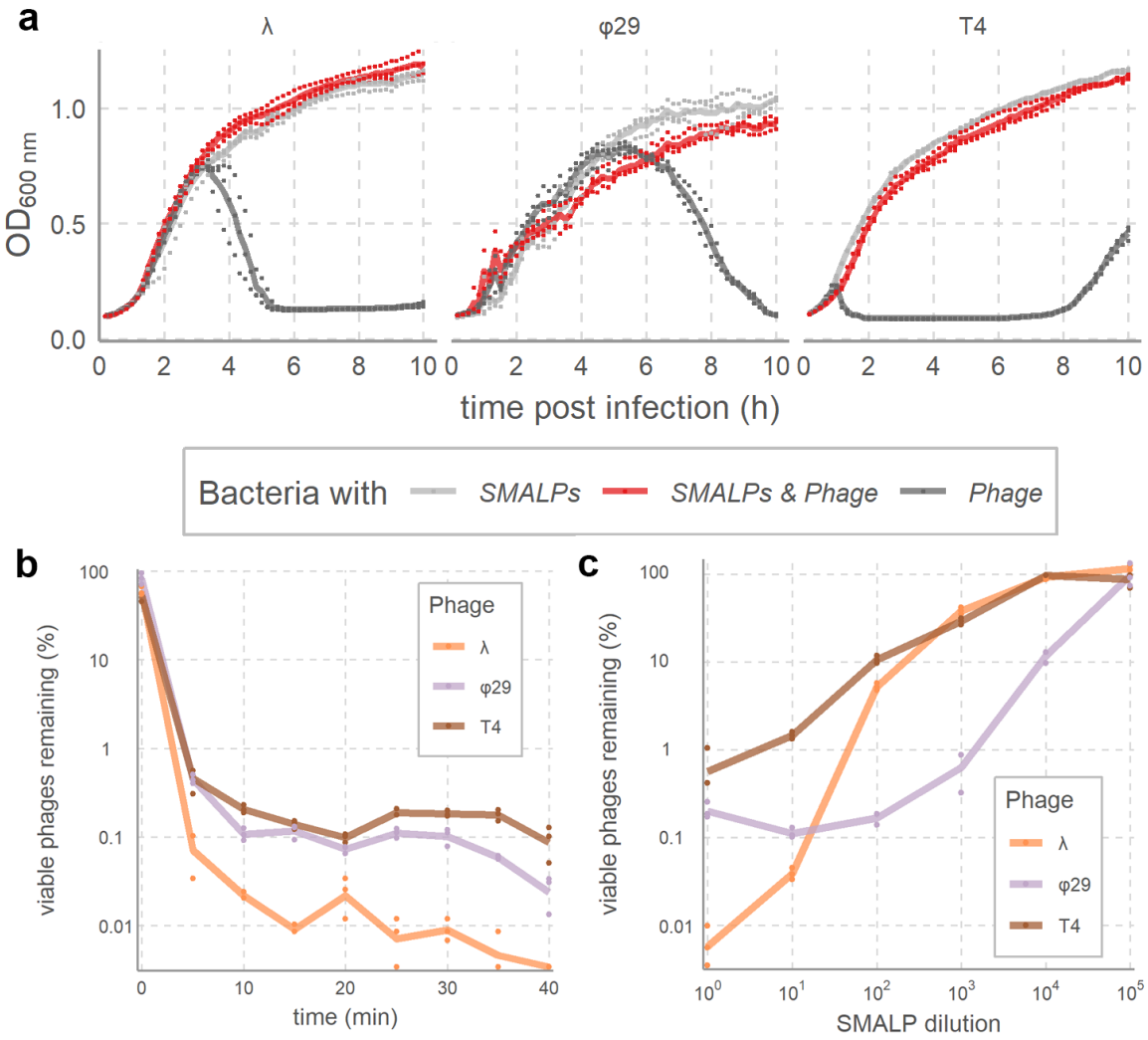


Figure 5: SMALPs inhibit phage lytic activity. (A) *Inhibition of lytic activity by Bacillus phage φ29 (10^6 pfu/ml, starting MOI: 10^{-2}) and Escherichia phages λ (10^5 pfu/ml, starting MOI: 10^{-3}) and T4 (10^4 pfu/ml, starting MOI: 10^{-4}) by SMALPs prepared from their hosts. Liquid cultures of host cells were either not infected (SMALPs-only control, light grey), infected with SMALPs-incubated phage (red), or infected with untreated phage (phage-only control, dark grey), followed by measurement of optical densities at 600 nm every 10 minutes. Points are measurements in each biological triplicate, solid lines are the mean.* ***(B)*** *Time-dependent inhibition of phages by SMALPs. Samples were taken at 5-minute time intervals of phages incubated with SMALPs from their host and used for plaque assays. The percentage of viable phages were estimated using a non-treated phage control. Points are replicates, while solid lines are means.* ***(C)*** *Phage sensitivity to SMALPs shown by treatments with increasingly*

diluted SMALPs. Experiment was carried out in the same manner as for (b), but with 20-minute time interval and tenfold dilutions of 10 mg/ml (original membrane concentration) SMALP stocks. Both (b) and (c) used 10^6 pfu/ml phage stocks.

Treatment of T4 at the same concentration with SMALPs prepared from an *E. coli* strain in which the outer sugar core of the LPS was removed (ΔRfaC), resulted in a slight decrease of spot size when compared to non-treated phage, as did *E. coli* ΔOmpC SMALPs. These results concur with research showing that T4 can efficiently infect *E. coli* strains with either one of its two receptors removed[334,342]. As treatment with these mutants resulted still resulted in some spot formation on the plates, these mutant trains caused incomplete inhibition of T4. This may result from the necessary presence of up to four separate interactions for successful T4 infection[341], as every T4 particle likely needs contact with multiple SMALPs to be completely inhibited. Negative stain transmission electron microscopy of phage T4 with *E. coli* K-12 SMALPs seemingly confirmed this, as it showed T4 tail fibers that each appeared to separately bind to SMALPs (Figure S1b-d). Interestingly, while some SMALPs seemed to interact with the distal tail tips, as is to be expected based on previous findings[343], other SMALPs could be interpreted as interacting further up the tail fibers. The presence of (presumably cellular) debris and the small size prohibited accurate quantifications of this effect. Therefore, it is uncertain if these observations result from imaging perspective. Future microscopy efforts could addressed such issues, , perhaps through employing covalently labelled SMALPs[344]. As SMALPs have already been used as platform to establish three-dimensional protein structures[324,345], these results indicated that SMALPs form an enticing tool to study the structure of phage particles bound to their receptor[316].

SMALPs cause receptor-specific genome ejection by phages.

To test whether phage-SMALP interactions led to DNA ejection, we measured fluorescence of DNA-specific fluorophores before and after incubation with SMALPs. This experiment employed the YO-PRO DNA stain, which was earlier shown to bind free DNA at a much faster rate than encapsulated phage DNA[346]. Additionally, this dye has been used to study DNA ejection in *Salmonella* phage P22[267]. We first tested our experimental setup by comparing fluorescence signal of λ phage stocks, *E. coli* SMALPs, and a combination of the two (Figure 3a). This revealed that SMALP stocks have sizeable fluorescence intensity when added to YO-PRO DNA stain. Fluorescence measurements of separate polymer confirmed that SMA itself

interacts with the DNA stain. This non-specific interaction with DNA stain likely results from the strong overall negative charge of the SMA polymer[347] making it similar to DNA. Despite this high background fluorescence signal, when we incubated SMALPs with phages, fluorescence significantly increased (two-tailed t-test, p = 0.0007, Figure 3a). Since the fluorescence of phage stocks was negligible, the increase evidently results from an increase of free DNA. Similar methods with the same DNA stain previously showed that fluorescence increase in *Salmonella* phage P22 in the presence of purified LPS is due to an increase in free DNA after being ejected by phage particles[267]. We therefore concluded that phage λ ejected its DNA in the presence of *E. coli* SMALPs.

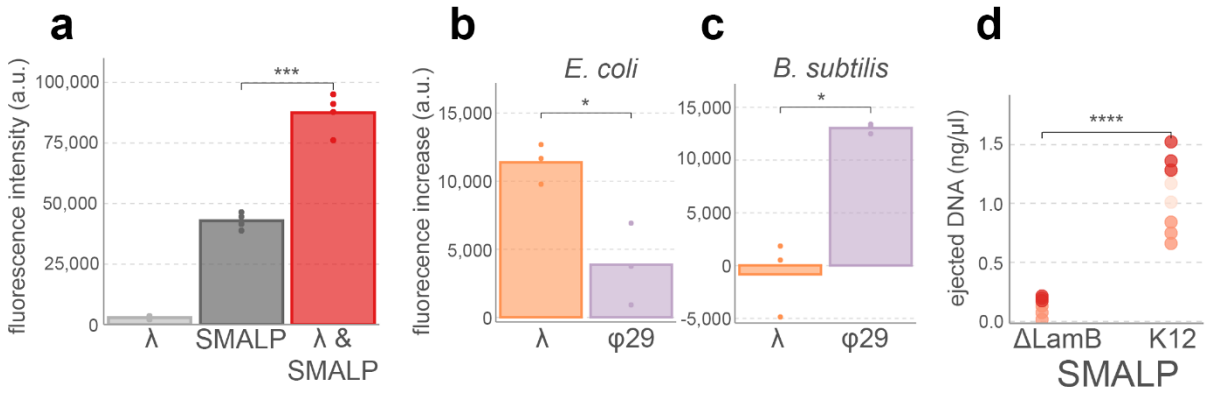


Figure 3: Specificity of ejection by various phages upon addition to SMALPs. (A) *Ejection of phage λ DNA in the presence of E. coli K12 SMALPs. YO-PRO fluorophore was added to phage λ stock (light grey), E. coli K-12 SMALPs (dark grey) and a combination of the two and fluorescence was measured after a 20-minute incubation. Points are biological replicates.* ***(B)*** *YO-PRO fluorescence assays like (a) performed with phages λ and φ29 added to E. coli K12 and B. subtilis SMALPs. SMALP-only measurements were subtracted from the SMALP + phage measurements, phage measurements were negligible (see (a)).* ***(C)*** *Same as b but using B. subtilis SMALPs.* ***(D)*** *Receptor specific DNA ejection by phage λ after incubation with SMALPs from E. coli K12 or ΔLamB, as measured by QuBit DNA quantification assays. The experiments were repeated thrice, each instance (which have different shades of red) composed of biological triplicates (individual points). SMALP-only measurements were subtracted from SMALP + phage incubations. Significance values are according to Welch two-tailed t-tests, * <0.05, ** <0.01, *** <0.001, **** <0.0001.*

Next, we examined whether phage DNA ejection was specific to SMALPs prepared using the appropriate host, we incubated phages λ and φ29 with SMALPs prepared from either *E. coli* or *B. subtilis*. After subtracting background signal from SMALP-only and phage-only samples, phage λ added to *E. coli* SMALPs had a significantly higher fluorescent signal than phage φ29 added to the same SMALPs (two-tailed t-test, $p = 0.03$, Figure 3b). Incubation of the two phages with SMALPs prepared from *B. subtilis* resulted in the reverse, with significantly higher background-adjusted fluorescence for phage φ29 (two-tailed t-test, $p = 0.02$, Figure 3c). From these results, we concluded that both phage λ and φ29 ejected their genomes in the presence of SMALPs prepared from their cognate hosts. Additionally, we concluded that SMALP-phage interactions reflect specific phage-host interactions, at least at this large phylogenetic distance (*E. coli* and *B. subtilis* are from different bacterial phyla and have different Gram stains).

We further determined the extent of specificity by comparing phage λ DNA ejection when added to SMALPs prepared from *E. coli* K-12 and *E. coli* ΔLamB. For these tests we employed a Qubit fluorometer, which is a calibrated system for DNA quantification. In three repeats of experiments that were each composed of biological triplicates, background-adjusted DNA measures were consistently higher when we added phage λ to *E. coli* K-12 SMALPs than to *E. coli* ΔLamB SMALPs (two-tailed t-test, $p = 6 \cdot 10^{-6}$, Figure 3d). Assuming that background-adjusted values were accurate representations of free DNA amounts, we estimated the fraction of phages which had ejected their DNA. Based on the stock phage titer of $1 \cdot 10^{10}$ pfu/ml and a molecular weight of $1.5 \cdot 10^{7}$ g/mol for λ DNA, the theoretical maximum of ejected DNA in our reaction was 0.80 ng/µl. The calibrated measurements identified 1.09 ± 0.28 ng/µl DNA in the reactions. Assuming a slight underestimation (i.e. well within an order of magnitude) of the added phage stock, these results indicated that (nearly) all phage particles ejected their genome. The SMALP-phage interaction is thus highly efficient at prompting DNA ejection in phage λ. Combined, these fluorescence-based experiments provided ample evidence that DNA ejection in SMALP-phage interactions is receptor specific. SMALPs could therefore be a powerful tool in studying phage-receptor interactions, especially for phages that interact with membrane proteins.

SMALPs as tool to study phage-host interactions.

While the above results show SMALPs are potentially useful tools to study phage-host interactions, they are also known to have drawbacks. One such drawback is that SMALPs are strong chelators of divalent cations[321]. This may be particularly problematic in phage-host

studies, as some phages are dependent on Mg^{2+} or other divalent cations for successful infection initiation[348,349]. Due to their chelating activity, proteins that are active against DNA, like polymerases, nucleases, or restriction enzymes, may not be active in the presence of SMALPs.

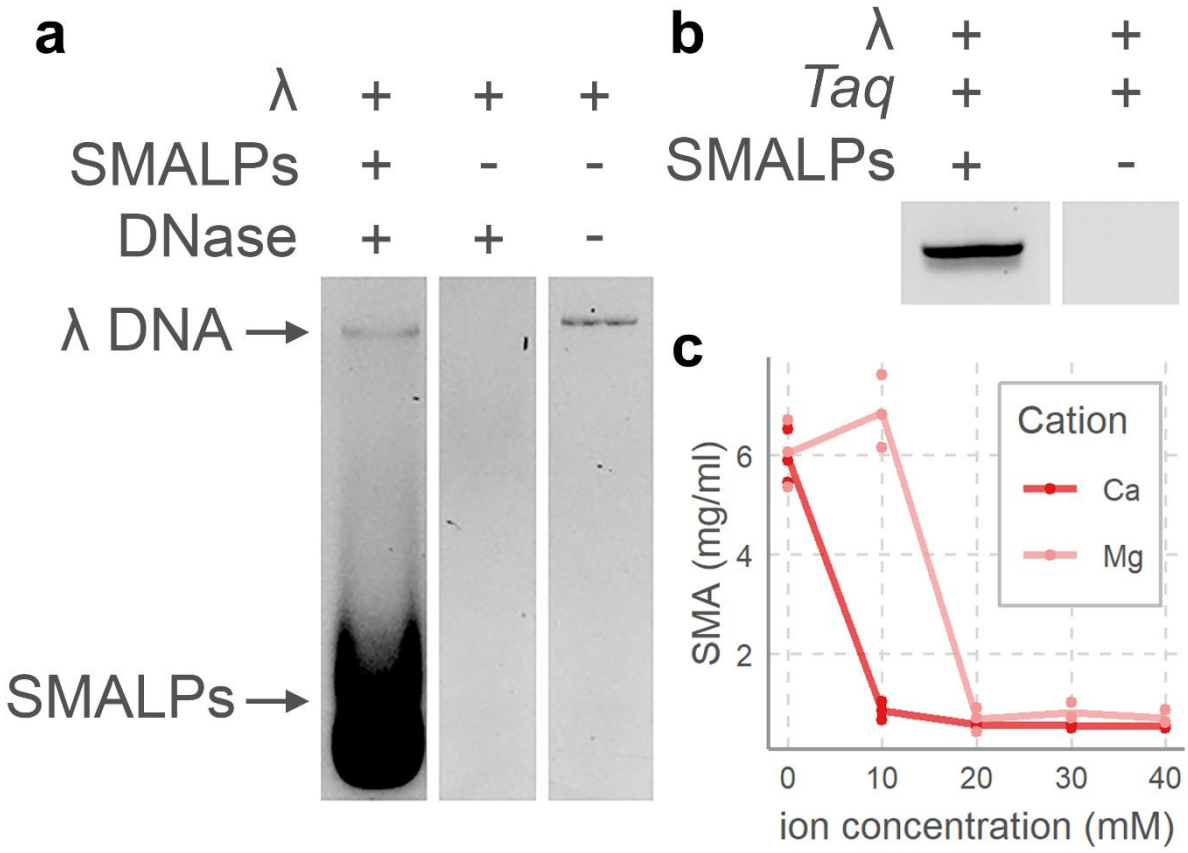


Figure 4: Chelating activity of SMALPs causes inhibition of DNA-modifying reactions. (A) *DNase inhibition by SMALPs, shown by an agarose gel of λ DNA with SMALPs and DNase, λ DNA with DNase, and λ DNA alone.* ***(B)*** *Inhibition of Taq polymerase by SMALPs after PCR targeting λ DNA.* ***(C)*** *Precipitation of SMA by addition of divalent cations in calcium chloride (Ca) and magnesium sulphate (Mg), followed by centrifugation. SMA concentration was determined by adsorption at 259 nm and comparison to a standard curve (Figure S2). Points are replicates, solid lines are averages.*

Indeed, when we tested the activity of DNase in the presence of SMALPs, we observed inhibition of nuclease activity (Figure 4a). This additionally supported the above findings that SMALPs bind to DNA-specific fluorophores, as they form thick smears at low molecular weight when subjected to agarose gel electrophoresis. Similarly, amplification reactions using *Taq* polymerase were entirely inhibited by SMALPs (Figure 4b). To circumvent this, we

developed a method of removing SMALPs from solution. This method is based on their chelating activity, which leads to SMALP precipitation[321]. Addition of calcium or magnesium to SMALP solutions above 10 mM successfully removed almost all polymer from solution (Figure 4c). However, SMALPs removal in this fashion could in turn inhibit polymerases, as high concentration of divalent cations are known to do[350,351]. Calcium particularly inhibits polymerases by outcompeting magnesium as co-factor[351], although high magnesium concentrations also inhibit polymerases[350,352]. Additionally, DNA molecules also interact with divalent cations[353], which should be taken into account for further studies of DNA released by phages in the presence of SMALPs. Applications of SMALPs to study phage-host interactions needs further development of alternative means of SMALP removal. Alternatively, several modified SMA polymers that are more positively charged and therefore potentially weaker chelators have recently been developed[344,354,355]. Their adoption for SMALP-phage studies may lift the above described drawbacks.

Conclusion

In this study, we report on the potential utility of SMALPs for studies of bacteriophage-receptor interactions. We found that SMALPs are viable tools, because their inactivation of phage particles results in host- and receptor-specific ejection of phage genomic material. As a result of our findings, SMALPs may see adaptation as a platform to study various aspects of phage-receptor interactions. SMALPs may be useful tool for cryo-electron microscopy of phages binding mechanisms[103,356–358], as this is implausible when using whole cells due to their size[359]. The differences in phage binding to 10 nm diameter SMALPs and to large bacteria, across the surface of which phages often engage in 2D diffusion[12], might further reveal details of the mechanics by which phages find their receptors. Additionally, affinity purification or pulldown assays of phage bound SMALPs coupled with mass spectrometry approaches may aid in receptor identification. While there are drawbacks associated with SMALPs (i.e. their chelation of magnesium), further developments of SMALPs as tools to study bacterial membrane proteins may aid in their future applications in studying phage-receptor interactions.

Methods

Bacterial and bacteriophage strains.

To test whether bacterial membranes from different bacteria could be solubilized by the SMA polymer, we used *E. coli* K-12 BW25113 (DSMZ #27469) and *Bacillus subtilis* 110NA (DSMZ #5547). To test specificity of phages to SMALPs, we further used *E. coli* strains JW3996 (ΔLamB), JW3596 (ΔRfaC), and JW2203 (ΔOmpC) from the Keio strain collection[269]. The phages used in this study were *Escherichia* phage λ (DSMZ #4499), *Bacillus* phage φ29 (DSMZ #5546) and *Escherichia* phage T4 (DSMZ #103876). All chemicals were obtained from Sigma Aldrich, except where stated otherwise.

Bacterial cultivation and phage production.

Bacteria were cultivated in Lysogeny Broth (LB) at 37°C under agitation at 180 rpm. Production of phage stocks was according to the soft-agar overlay method as described before[360]. Phage stocks were kept at 4°C until further use. To enumerate phages, 0.1 ml exponentially growing bacterial culture was added to 5 mL 0.7% (w/v) LB agarose (Bio-Rad), which was subsequently layered on a 1.5% (w/v) LB agar plate. After the top layer had dried, 10µl of 10-fold dilution ranges of phage stocks in SM buffer (100 mM NaCl, 8 mM $MgSO_4$ x 7 H_2O, 50 mM Tris-HCl pH 7.5) were pipetted on the agar. The plates were placed at a 45° angle and until the phage dilutions had dried and then incubated overnight at 37°C).

Styrene Maleic anhydride co-polymer hydrolyzation.

Styrene maleic anhydride co-polymer (67:33 ratio, SManh) (Polyscience) was hydrolyzed to styrene maleic acid co-polymer as described before[321]. In short, 12.5 g of SManh was suspended in a round bottom flask containing 250 ml 1M NaOH. This flask was connected to a reflux setup and the suspension was heated in sunflower oil at 98°C for 4 hours under constant stirring. Afterward, the suspension was cooled to room temperature and 6 M HCl was added to a final concentration of 1.1 M to precipitate the hydrolyzed polymer. Styrene maleic acid (SMA) was pelleted by centrifugation at 7000 x *g* for 20 minutes, after which the supernatant was discarded. The pellet was resuspended in 250 mL 100 mM HCl and centrifuged at 7000 x *g* for 20 minutes, after which the supernatant was discarded. This was repeated twice, once with 250 mL 100 mM HCl and once with deionized water. The SMA pellet was frozen at -80°C and extensively freeze dried. SMA was dissolved at a concentration of 60 mg/ml in 20 mM Tris-HCl (pH 8) and stored at -20°C until further use.

Bacterial membrane isolation and SMALP production.

To isolate bacterial cell membranes, bacteria were first grown overnight at 37°C while shaking. Cells were pelleted by centrifugation at 10,000 x *g*, 4°C for 15 minutes, after which supernatant was discarded. Pellets were washed with the original volume of lysis buffer (50 mM Tris HCl pH 7.5, 2 mM $MgCl_2$), after which centrifugation was repeated. After discarding the supernatant, cells were re-suspended in 4 volumes of lysis buffer per gram wet cell weight, to which 1 tablet of cOmplete EDTA-free protease inhibitor was added. To lyse the cells, the suspension was thrice passing through a model CF1 Cell Disruptor (Constant Systems) at 1.5 kBar. Cell debris was pelleted by centrifugation at 12,000 x *g*, 4°C for 15 minutes and collecting the supernatant. Next, membranes were pelleted by centrifuging the supernatants at 225,000 x *g* ,4°C for 1 hour. Pellets were re-suspended in lysis buffer and centrifuged again. Soluble proteins were removed by resuspending the pellet in 200 mM NaCl + 20 mM Tris HCl, pH 7.5 and repeating centrifugation. Pellets were dissolved in 20 mM Tris HCl pH 7.5 + 200 mM NaCl at a concentration of 10 mg/ml and stored at -20°C until further use.

To produce SMALPs from bacterial membranes, we suspended membranes in 200 mM NaCl + 20 mM Tris HCl, pH 7.5 and added SMA to obtain a final solution with 10 mg/ml membrane and 20 mg/ml SMA (or a membrane:polymer ratio of 1:2). To allow SMALP formation, solutions were incubated under constant rotation for 20 minutes at room temperature. Non-solubilized membrane material and excess SMA were removed by filtration through a sterile 20 µm filter and extensive dialyzing against 20 mM Tris HCl pH 7.5. SMALP solutions were stored at 4°C until further use.

The effects of divalent cations on the SMA polymer were examined using $CaCl_2$ and $MgSO_4$. To 14 mg/ml SMA polymer, 0-30 mM $CaCl_2$ or $MgSO_4$ were added, and the $OD_{259\ nm}$ was measured in a cuvette using a NanoPhotometer C40 (Implen). The concentration of SMA in every sample was calculated using a linear standard curve ranging from 0-40 mg/ml SMA.

To determine SMALP sizes, we used dynamic light scattering using a Zetasizer ZS instrument (Malvern), using default software settings and multiple narrow modes analysis of the correlation data. Before measurement, samples were briefly degassed and equilibrated for 300 s at room temperature.

To test the efficacy of SMALP dissolution for different bacterial membranes, membrane suspensions of 10 mg/ml were prepared in 200 mM NaCl + 20 mM Tris HCl, pH 7.5 in 1.5 ml cuvettes. Subsequently, SMA was added to a final concentration of 20 mg/ml and the $OD_{600\ nm}$ was followed over time in a NanoPhotometer C40 (Implen), with measurements every 10 s.

Effect of SMALPs on phage lytic activity.

Equal volumes of SMALP solution and phage stock at 10-fold dilutions between 10^1 and 10^{10} pfu/ml were mixed and incubated for 20 minutes at room temperature to allow phages to bind SMALPs. SMALP and phage dilutions were prepared in sterile dilution buffer (100 mM NaCl + 50 mM Tris HCl pH 7.5) where necessary. Bacteria were grown to an $OD_{600\ nm}$ of 0.5 and 10 µl bacterial suspension was diluted in 89 µl fresh LB medium in a 96 wells plate. Diluted bacteria were incubated at 37°C under agitation for 30 minutes, after which 1 µl of SMALP/phage suspension was added. To bacteria-only controls, 1 µl dilution buffer was added instead. The 96-wells plate was then incubated in a Synergy H1 microplate reader (Biotek) at 37°C under continuous double orbital shaking for 10 hours, during which the $OD_{600\ nm}$ was determined every 10 minutes.

Characterizing SMALP-phage interaction with agar plate assays.

To determine the effect of incubation length on SMALP-phage interactions, equal volume of phage (at 10^6 pfu/ml) and SMALP solutions were mixed. At five-minute time intervals between 0 and 40 minutes, 20 µl aliquots were retrieved, to which $CaCl_2$ was added to a final concentration of 20 mM. As negative control to establish original viable phage counts, phage stock at 10^6 pfu/ml was diluted in an equal volume of 200 mM NaCl + 20 mM Tris HCl, pH 7.5, to which $CaCl_2$ was added to a final concentration of 20 mM. After a two-minute centrifugation at 21,000 x *g*, the supernatant was retrieved and used to enumerate viable phage particles as described under "Bacterial cultivation and phage production". To test the effect of SMALP concentration on phage inhibition, 10-fold SMALP dilutions were made in 100 mM NaCl + 50 mM Tris HCl pH 7.5. Samples were subsequently prepared and enumerated in similar fashion as in the time trials, except that only incubation times of 20 minutes were used. Negative controls were produced in the same manner as for the time trials. Decreases in viable phage particles were calculated by dividing phage titers after the reaction by those obtained from negative controls. For both time trial and SMALP concentration experiments, all samples consisted of biological triplicates.

Spot assays to determine specificity.

SMALP and phage reactions, as well as negative controls were performed as described in the previous section, with incubation times of 20 minutes. After incubation, 10 µl spots of reaction mixture were placed on top of a double layer agar plate prepared as described under "Bacterial

cultivation and phage production". Spots were dried at room temperature and subsequently incubated for 16 hours at 37°C.

Transmission electron microscopy.

Before transmission electron microscopy, phage stocks were purified using a preformed CsCl density gradient. A two-step gradient was prepared in thin walled ultracentrifuge tubes (Beckman Coulter) using CsCl at densities of 1.6 g/ml and 1.4 g/ml. On top of these layers, 1 ml 10^{11} phage preparation was placed, and density gradients were centrifuged at 111,000 x *g* for 2 hours using a Sw60Ti swinging bucket rotor (Beckman Coulter). After centrifugation, white phage layers were collected by puncturing the tubes with a hypodermic needle, as was advised before[361]. To remove excess CsCl, phages were cleaned by three consecutive washes with SM-buffer in an Amicon 30 kDa spin filter (Merck), which was centrifuged for 10 minutes at 3,000 x *g*. Samples with SMALPs were then prepared as described under "Characterizing SMALP-phage interaction with agar plate assays" with an incubation time of 20 minutes and phage titers of 10^{10} pfu/ml. Samples were applied to thin carbon-coated 400 square mesh copper grids (Electron Microscopy Sciences), which were ionized by glow discharged 90 seconds. On top of the grids, 3 µl sample were carefully pipetted and incubated at room temperature for 1 minute. Liquid was removed using filter paper (Whatman). Grids were then washed thrice with 10 µl milli-Q water, each time removing liquid with filter paper. Finally, 3 µl 2% uranyl acetate was pipetted on the grids. After a final 30 s incubation at room temperatures, uranyl acetate was removed with filter paper. TEM imaging used a Philips CM200 (200 kV), while micrographs were captured using a TemCam- F416 4 kkD (TVIPS) at 150,000× magnification using EM-MENU software.

Fluorescence DNA ejection assays.

The first fluorescence assay tested DNA ejection with *Escherichia* phage λ and SMALPs prepared from *E. coli* K12 BW25113 membranes. Phage stock was diluted to 10^6 pfu/ml using SM buffer. Phage stock, SMALP stock, and the two combined were incubated with 1.1 µM YO-PRO-1 iodine (491/509, Invitrogen) for 20 minutes at 37°C, as described before for P22 and *S. enterica* LPS[267]. Fluorescence was then measured using a QuBit 4 fluorometer (ThermoFisher), which used an excitation wavelength of 430-495 nm and measured emission at 510-580 nm. Samples were composed of biological triplicates. To determine specificity, *Escherichia* phage λ, *Bacillus* phage φ29, *E. coli* K12 BW25113 SMALPs, and *B. subtilis*

110NA SMALPs were measured separately and in every possible combination of phage and SMALP in the same way.

For further quantification of DNA ejection after SMALP concentration, the assay was repeated with $1 \cdot 10^{10}$ pfu/ml *Escherichia* phage λ, and SMALPs from *E. coli* K12 BW25113 and ΔLamB, using a dsDNA HS assay kit and Qubit fluorometer (Thermo Fisher Scientific). This assay consisted of three repeats, with each repeat consisting of biological triplicates. We calculated the theoretical amount of DNA that was present in the phages in the sample using:

$$[\mathrm{DNA}] = \frac{M_{DNA}}{A} \cdot 10^6[\varphi]$$

Where [DNA] is the DNA concentration in ng/µl, M_{DNA} is the molecular weight of λ DNA ($1.5 \cdot 10^7$ pfu/ml), A is Avogadro's number, and φ is the phage titer in the reaction ($3.2 \cdot 10^{10}$ pfu/ml).

SMALP assays with divalent cations.

To determine the effect of SMALPs on DNase activity, we prepared SMALPs from bacterial membranes as described above. Three samples were made for the assay, the first containing 9 µl SMALP solution, 10 ng phage λ DNA (New England Biolabs), and 2 U DNase I. The second sample replaced the SMALPs with deionized water, and the third sample replaces both the SMALPs and DNase I with water. The samples were incubated at 37°C for 20 minutes and ran on a 1% (w/v) agarose gel for 30 minutes at 20 V/cm.

To test inhibition of *Taq* polymerase by SMALPs, PCR reactions were made containing 1x *Taq* polymerase Master Mix (NEB), 0.2 µM forward (TACGCCGGGATATGTCAAGC) and reverse primers (TACGCCAGTTGTACGGACAC) that target the phage λ E gene, 0.1 ng phage λ DNA. In one sample, 10x diluted SMALP solution was added. PCR program consisted of 30 seconds at 95°C, 25 cycles of 30 seconds at 95°C, 30 seconds at 55°C, and 60 seconds at 72°C, and 5 minutes at 72°C. Samples were then run on agarose gel as described above.

The effect of the divalent cations Ca^{2+} and Mg^{2+} was tested as follows. SMA stocks of 6 mg/ml were incubated for 20 minutes in the presence of 0-40 mM of either $CaCl_2$ (Sigma) or $MgCl_2$. SMA precipitate was pelleted by centrifuging for 1 minute at 21,000 x *g*. Next, adsorption of the supernatant was determined at 259 nm on a NanoPhotometer C40 (Implen). Concentrations were calculated using an SMA standard curve with concentrations ranging from 0-40 mg/ml.

Supplementary Figures

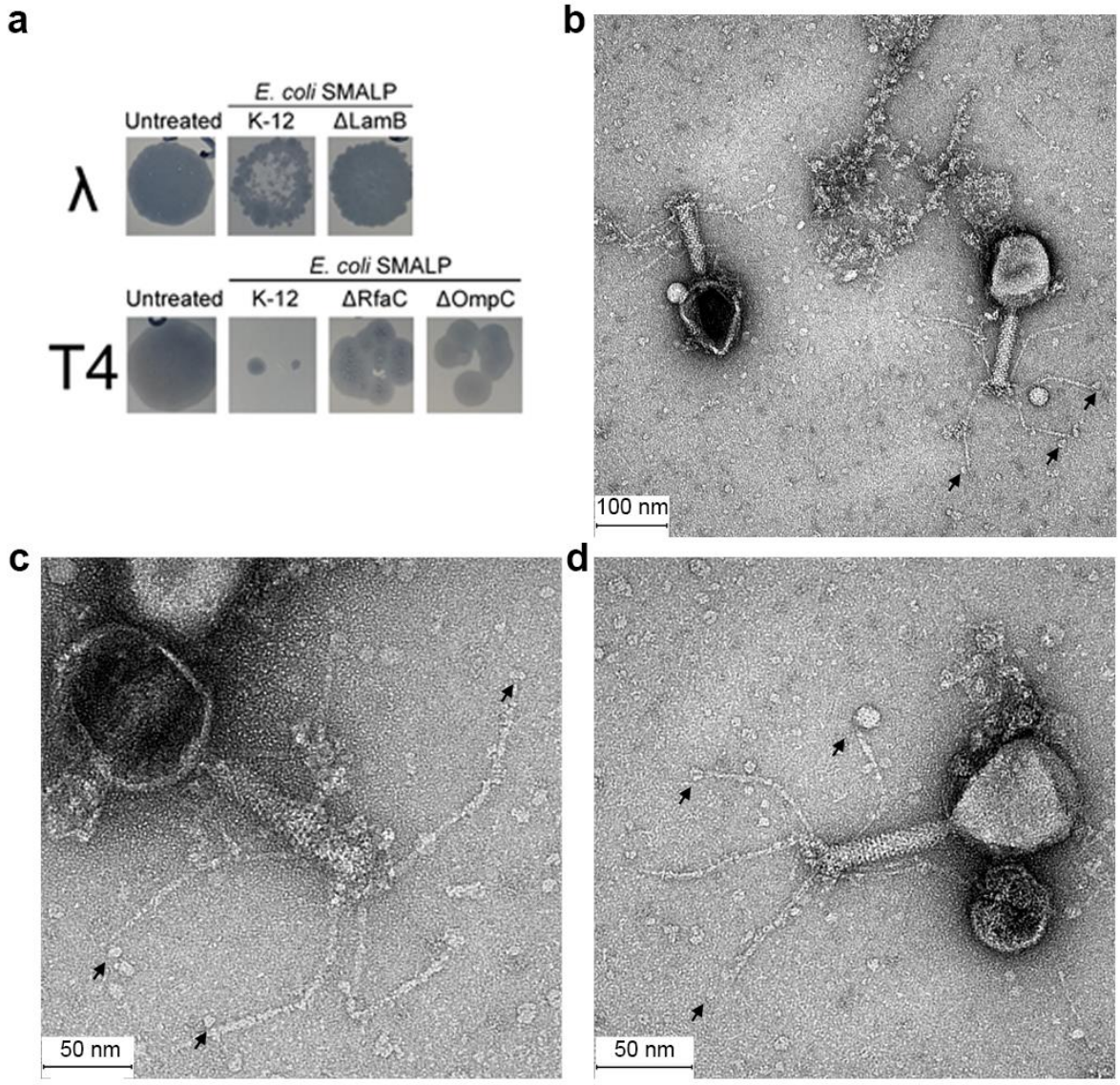


Figure S1: Characterization of the SMALP-phage interaction. (A) *Specificity of the SMALP-phage interaction for λ and T4 indicated by spot assays of untreated phages, phages treated with E. coli K12 SMALPs, and phages treated with SMALPs from which phages receptors were removed. E. coli ΔRfaC lacked the LPS receptor of T4, ΔOmpC lacked the outer membrane protein C receptor of T4, ΔLamB lacked the maltose porin protein receptor of λ. Phage titers were 10^6 pfu/ml for phage λ and 10^5 pfu/ml for phage T4.* ***(B)-(D)*** *Micrographs of phage T4 show phage particles in the vicinity of multiple independent SMALPs. Arrows point to SMALPs. Micrographs were captured at magnifications of 60,000x (b), 100,000x (c) and 150,000x (d), while scale bars denote 100 nm (b), 50 nm (c) and 50 nm (d).*

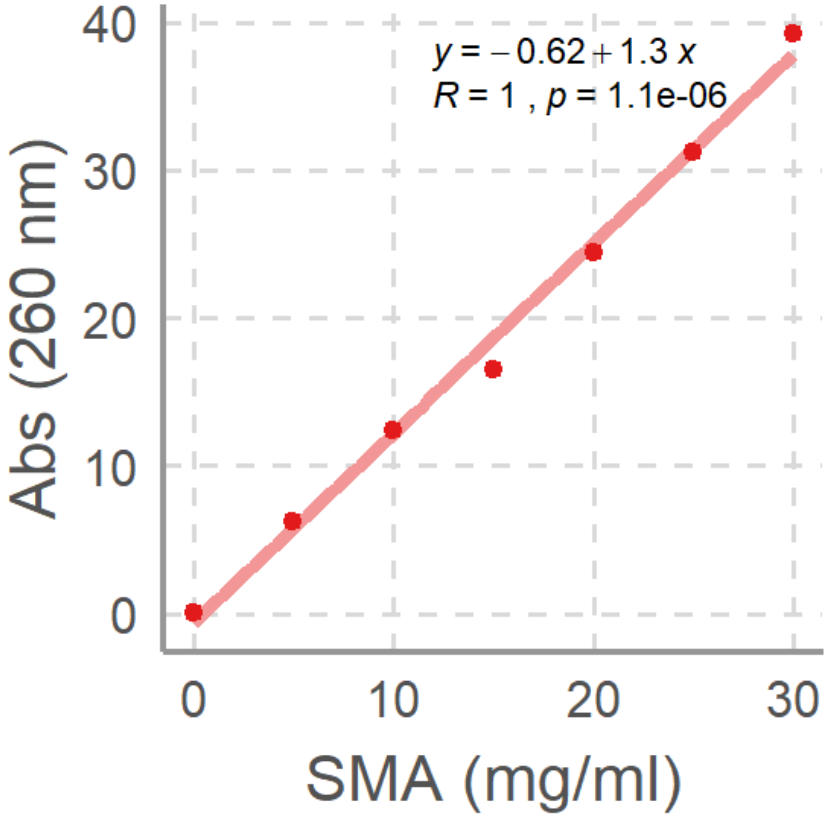


Figure S2: Standard curve used to calculate SMA concentrations in Figure 4c.

6

Concluding remarks and outlook

"For even the wise cannot see all ends."

The Fellowship of the Ring; J.R.R. Tolkien

To fully understand the impact of phages on the biosphere, it is necessary to study the composition of viral dark matter and the molecular interactions between phages and bacteria. These then were the pervading themes of the research presented in this thesis. In this final chapter, final remarks on the previous four chapters will be provided.

On studying viral dark matter

As described in chapter 1, most phages remain unknown. Uncovering and understanding this viral dark matter is arguably the main challenge and focus of modern phage biology. Chapters 2 and 4 present research aiding in this endeavour.

Chapter 2 presents an in-depth analysis of BACON domains in crAss-like phages, thereby advancing understanding of viral domain tandem repeat evolution. In addition, this study expanded known phage biodiversity by uncovering two distinct phage lineages with crAssBACON domains. The first of these, which we referred to as cluster A, is a distant relative of the crAss-like phages. The lineage carries crAss-like terminase and portal proteins, but lacks identifiable homologs to the other structural proteins with which crAss-like phage phylogenies have been defined[86]. While our analyses in chapter 2 did not aim to comprehensively define a new phage lineage, it established cluster A phages as a distinct phage lineage. Future expansion of cluster A and analysis of its relation to crAss-like phages may assign it to a separate phage (sub-)family in the context of broader crAss-like phage taxonomy. Further differentiating it from crAss-like phages is the apparent temperate nature of cluster A phages, based on the presence of antirepressor and transposase genes in cluster A genome. If these phages were classified within the crAss-like phage lineage, they would thereby form the first examples of temperate crAss-like phages.

The second new lineage with crAssBACON domains, cluster B, is in itself enigmatic. Based on the high number of genes homologous to *Acinetobacter* and *Klebsiella* in cluster B genomes, this lineage seems to represent phages infecting members of the *Gammaproteobacteria*. This lineage points to two interesting open questions. The first question is whether these are phages with a broad host-range that infect across orders in the *Gammaproteobacteria*, or (more likely) multiple phage species that infect a wide range of *Gammaproteobacteria.* In both cases, there is the second question of how these phages obtained their crAssBACON domains. The obvious answer to this is horizontal gene transfer, but the exact path by which crAssBACON domains transferred between phages that likely infect *Proteobacteria* and *Bacteroidetes* remains unclear.

Taken together, the study of crAssBACON domains showed how the analysis of a single protein domain can aid in the study of uncultivated phage lineages. Modern phage metagenomics, particularly in the human gut, tends to focus on sweeping community-wide analyses. While these analyses are providing needed insights into the role of phages in human health, chapter 2 shows how studies focusing on single traits provide a complementary path. Such studies, focusing on a single trait and/or a single phage lineage, may contextualize the role of phages within the complexities of intestinal phage community dynamics. Furthermore, adding studies of single genes, particularly if a function is described, to the databases of phage knowledge helps to shed light on dark matter. This is because detection of homology between an unknown phage gene and genes encoding a known function is a main method by which insights and novel research avenues are created. As such, studies such as that presented in chapter 2 can be an interesting approach in phage biology.

Besides directly reporting novel phage lineages using crAssBACON domains, chapter 4 presents AdsorpSeq as a method to assign phage genomes from environmental viromes to a bacterial host. After vetting the technique on model phages, we applied it to a hospital wastewater virome and showed that we could assign hosts to uncharacterised phage sequences. Like any method, AdsorpSeq has its limitations. Perhaps premier among these is its dependency on bacterial cultivation. As explained in chapter 1, this excludes a large number of bacteria that either have not yet been or cannot be isolated in a laboratory setting. While thus limited, compared to classic isolation techniques AdsorpSeq still increases the number of bacteria for which phages can be studied due to its use of liquid cultivation. Such cultivation particularly simplifies studies of phages that infect anaerobic bacteria (e.g. gut *Bacteroides*), as in AdsorpSeq these can be cultivated in closed vessels without the need of special equipment like anaerobic tents. The culture-dependence of AdsorpSeq could also be a strength because it allows targeted studies. For example, it may be useful to study which phages from an environment interact with a specific bacterial species. This could be useful for analyses of phage-host interactions in an environment over time, or when comparing the phages that interact with a host species between multiple environments. While such studies are possible with traditional methods like plaque assays, unlike these methods AdsorpSeq does not suffer from selection biases toward narrow host-range lytic phages (chapter 3). It thus allows the characterisation of all phages in the environment regardless of host range. Besides species, analysis of phages that interact with a specific receptor, like an antibiotic efflux pump, can be performed using AdsorpSeq. Through such targeted studies AdsorpSeq fills a niche in studies of viral dark matter.

On understanding phage host-interactions

Besides uncovering novel phages, the work I presented in this thesis endeavoured to better understand the interactions between phages and hosts. We analysed BACON domains (chapter 2) because they were originally hypothesised to be directly involved in host interactions, although in light of current knowledge they seem not to engage in carbohydrate binding. In chapter 3, we presented a novel framework for understanding phage host range and provided current knowledge on the mechanisms regulating it. Finally, in chapter 5 we explored the potential to use SMALPs as a tool to study phage-host interactions. Particularly in the latter two chapters phage-host interactions were the focus.

The main novelty of the framework that we provided in chapter 3 was in distinguishing between the host range of a lineage/species and that of a particular phage particle. This was absent from existing literature, possibly leading to confusion on the meaning of terms such as narrow versus broad host-range. For example, it has recently been reported that narrow host-range seems highly prevalent among gut phages[88], a conclusion that our results in chapter 4 agreed to. However, in the light of the framework presented in chapter 3, the highly specific phage particles represented in these studies may be part of phage quasispecies with broader host-range. This may also be true, as argued above, in cluster B from chapter 2. Proof for such hypotheses is of course dependent on finding highly related phages with different host ranges, for which the methods presented in chapter 4 could serve. Furthermore, one could argue that according to this model ***all*** phage lineages have a broad host range at some taxonomic level. In fact, this hypothesis should perhaps be the focus of separate research.

In the end, it is important to distinguish between such broad host-range quasispecies and true broad host range particles that can infect multiple hosts. This is because there are separate potential implications and applications of the phage tropism switching that occurs in broad host range species and the complex machinery required for a polyvalent phage particle. The former might be common and must be studied thoroughly to allow phage applications like phage therapy. Meanwhile, the latter are of interest as molecular machines. Beyond the ecological and evolutionary implications of broad host-range, much can be learnt from the machinery that has evolved to allow one particle to adhere to and infect two or more types of host.

In order to study such machinery, SMALPs could be highly useful. As SMALPs allow for the isolation of single host receptors without disturbing their natural environments, they could allow microscopy studies of broad host-range phage particles as they are bound to their receptor. Such studies arguably form the biggest potential for SMALPs as a method to study

phage-host interactions. Of course, as alluded to in chapter 5, SMALPs have their drawbacks. These are mainly related to their chelation of divalent cations, which disturb DNA editing reactions (e.g. amplifications) and other biological reactions[321]. Such issues are also related to the novel nature of the SMALPs method, which is still under active development (see e.g. [330,362,363]). Future applications of SMALPs could thus need extended development of the SMALP polymer, such as those that potentially alter their chelating behaviour (e.g. [355]).

The future of phage research

The time frame at which the research presented in this thesis was performed marked the centennial of phage research, both when counting from the discoveries of Frederick Twort in 1915 and when counting from those of Felix d'Herelle in 1917[38,229,364]. Around this centennial, phage biology is undergoing a renaissance. Societally, there is a noticeable recent uptick in interest in phages (e.g. [365–367]) almost exclusively concerning phage therapy as a potential alternative to antibiotics treatment. Academically, the past few years have seen major new discoveries in phage biology. These included several entirely novel mechanisms encoded by phages, such as phage nuclei[200,292] and tubulins[368], phage quorum sensing[369,370], and phage CRISPR-Cas systems[371]. Phages have historically served as a conduit for understanding cellular life (e.g. in understanding the function of DNA) and molecular biology highly depends on phage-derived proteins (e.g. polymerases). As such, I feel that the above recent discoveries are a continuation to this trend, and I hope that the frameworks and tools provided in this thesis will aid in uncovering additional unique phage-encoded mechanisms. Next to novel mechanisms, the major breakthroughs in recent phage biology concern our increasing understanding of phage ecological dynamics and biodiversity. This has included the description of novel mechanisms regulating phage ecology[11,160,372] and the description of entire new phage lineages[83,85,86]. Here, more directly, the work presented in chapters 3, 4, and 5 will aid in future endeavours.

References

1. Moreira, D. & López-García, P. Ten reasons to exclude viruses from the tree of life. *Nat. Rev. Microbiol.* **7**, 306–311 (2009).
2. Claverie, J. M. & Abergel, C. Mimivirus: The emerging paradox of quasi-autonomous viruses. *Trends Genet.* **26**, 431–437 (2010).
3. Forterre, P. To be or not to be alive: How recent discoveries challenge the traditional definitions of viruses and life. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 100–108 (2016).
4. Raoult, D. & Forterre, P. Redefining viruses: Lessons from Mimivirus. *Nat. Rev. Microbiol.* **6**, 315–319 (2008).
5. Koonin, E. V. & Starokadomskyy, P. Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. *Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci.* **59**, 125–134 (2016).
6. Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene. *Nature* **260**, 500–507 (1976).
7. Devoto, A. E. *et al.* Megaphages infect Prevotella and variants are widespread in gut microbiomes. *Nat. Microbiol.* **4**, 693–700 (2019).
8. Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425–431 (2020).
9. Zandi, R., Reguera, D., Bruinsma, R. F., Gelbart, W. M. & Rudnick, J. Origin of icosahedral symmetry in viruses. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15556–15560 (2004).
10. Jazwinski, S. M., Lindberg, A. A. & Kornberg, A. The gene H spike protein of bacteriophages ØX174 and S13. I. Functions in phage-receptor recognition and in transfection. *Virology* **66**, 283–293 (1975).
11. Barr, J. J. *et al.* Bacteriophage adhering to mucus provide a non-host-derived immunity. *Proc. Natl. Acad. Sci.* **110**, 10771–10776 (2013).
12. Nobrega, F. L. *et al.* Targeting mechanisms of tailed bacteriophages. *Nat. Rev. Microbiol.* **16**, 760–773 (2018).
13. Ackermann, H.-W. Phage Classification and Characterization. in *Bacteriophages: Methods in Molecular Biology* (eds. Clokie, M. R. & Kropinski, A. M.) **30**, 127–140 (Humana Press, 2009).

14. Fokine, A. *et al.* A three-dimensional cryo-electron microscopy structure of the bacteriophage φKZ head. *J. Mol. Biol.* **352**, 117–124 (2005).
15. Catalão, M. J., Gil, F., Moniz-Pereira, J., São-José, C. & Pimentel, M. Diversity in bacterial lysis systems: Bacteriophages Show the way. *FEMS Microbiol. Rev.* **37**, 554–571 (2013).
16. Zhao, Y. *et al.* Searching for a 'Hidden' Prophase in a marine bacterium. *Appl. Environ. Microbiol.* **76**, 589–595 (2010).
17. Weinbauer, M. G. Ecology of prokaryotic viruses. *FEMS Microbiol. Rev.* **28**, 127–181 (2004).
18. Howard-Varona, C., Hargreaves, K. R., Abedon, S. T. & Sullivan, M. B. Lysogeny in nature: Mechanisms, impact and ecology of temperate phages. *ISME J.* **11**, 1511–1520 (2017).
19. Christie, G. E. & Dokland, T. Pirates of the Caudovirales. *Virology* **434**, 210–221 (2012).
20. Erez, Z. *et al.* Communication between viruses guides lysis–lysogeny decisions. *Nature* **541**, 488–493 (2017).
21. Trinh, J. T., Székely, T., Shao, Q., Balázsi, G. & Zeng, L. Cell fate decisions emerge as phages cooperate or compete inside their host. *Nat. Commun.* **8**, 14341 (2017).
22. Łoś, M. & Węgrzyn, G. Pseudolysogeny. in *Advances in Virus Research* (eds. Łobocka, M. & Szybalski, W.) **82**, 339–349 (Academic Press, 2012).
23. Barr, J. J. A bacteriophages journey through the human body. *Immunol. Rev.* **279**, 106–122 (2017).
24. Hannigan, G. D. *et al.* The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome. *MBio* **6**, e01578-15 (2015).
25. Lim, Y. W. *et al.* Metagenomics and metatranscriptomics: Windows on CF-associated viral and microbial communities. *J. Cyst. Fibros.* **12**, 154–164 (2013).
26. Dickson, R. P. & Huffnagle, G. B. The Lung Microbiome: New Principles for Respiratory Bacteriology in Health and Disease. *PLoS Pathog.* **11**, 1–5 (2015).
27. Abeles, S. R. *et al.* Human oral viruses are personal, persistent and gender-consistent. *ISME J.* **8**, 1753–1767 (2014).
28. Santiago-Rodriguez, T. M., Ly, M., Bonilla, N. & Pride, D. T. The human urine virome in association with urinary tract infections. *Front. Microbiol.* **6**, 1–12 (2015).
29. Rodela, M. L., Sabet, S., Peterson, A. & Dillon, J. G. Broad environmental tolerance for a salicola host-phage pair isolated from the cargill solar saltworks, newark, ca, usa.

Microorganisms **7**, (2019).

30. Zhang, X. & Wang, Y. Genome analysis of deep-sea thermophilic phage D6E. *Appl. Environ. Microbiol.* **76**, 7861–7866 (2010).
31. Filippova, S. N. *et al.* Detection of phage infection in the bacterial population of Lake Untersee (Antarctica). *Microbiol. (Russian Fed.* **82**, 383–386 (2013).
32. Kim, M. S. & Bae, J. W. Lysogeny is prevalent and widely distributed in the murine gut microbiota. *ISME J.* **12**, 1127–1141 (2018).
33. Touchon, M., Bernheim, A. & Rocha, E. P. C. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.* **10**, 2744–2754 (2016).
34. Silveira, C. B. & Rohwer, F. L. Piggyback-the-Winner in host-associated microbial communities. *npj Biofilms Microbiomes* **2**, 16010 (2016).
35. Knowles, B. *et al.* Variability and host density independence in inductions-based estimates of environmental lysogeny. *Nat. Microbiol.* **2**, 17064 (2017).
36. Cobián Güemes, A. G. *et al.* Viruses as Winners in the Game of Life. *Annu. Rev. Virol.* **3**, 197–214 (2016).
37. Mokili, J. L., Rohwer, F. & Dutilh, B. E. Metagenomics and future perspectives in virus discovery. *Curr. Opin. Virol.* **2**, 63–77 (2012).
38. Salmond, G. P. C. C. & Fineran, P. C. A century of the phage: past, present and future. *Nat. Rev. Microbiol.* **13**, 777–786 (2015).
39. Bergh, Ø. *et al.* High abundance of viruses found in aquatic environments. *Nature* **340**, 467–8 (1989).
40. Marov, M. Y. The Structure of the Universe. in *The Fundamentals of Modern Astrophysics: A Survey of the Cosmos from the Home Planet to Space Frontiers* 279–294 (Springer New York, 2015). doi:10.1007/978-1-4614-8730-2_10
41. Brum, J. R. *et al.* Ocean Viral Communities. *Science (80-.).* **348**, 1261498-1–11 (2015).
42. Youle, M., Haynes, M. & Rohwer, F. Scratching the Surface of Biology's Dark Matter. in *Viruses: Essential Agents of Life* (ed. Witzany, G.) 61–81 (Springer Netherlands, 2012). doi:10.1007/978-94-007-4899-6_4
43. Puxty, R. J., Millard, A. D., Evans, D. J. & Scanlan, D. J. Viruses Inhibit CO2 Fixation in the Most Abundant Phototrophs on Earth. *Curr. Biol.* **26**, 1585–1589 (2016).
44. Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* **438**, 86–9 (2005).
45. Fridman, S. *et al.* A myovirus encoding both photosystem i and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. *Nat. Microbiol.* **2**, 1350–1357

(2017).

46. Lu, M. J. & Henning, U. Superinfection exclusion by T-even-type coliphages. *Trends Microbiol.* **2**, 137–9 (1994).
47. Bondy-Denomy, J. & Davidson, A. R. When a virus is not a parasite: the beneficial effects of prophages on bacterial fitness. *J. Microbiol.* **52**, 235–242 (2014).
48. Vica Pacheco, S., García González, O. & Paniagua Contreras, G. L. The lom gene of bacteriophage lambda is involved in Escherichia coli K12 adhesion to human buccal epithelial cells. *FEMS Microbiol. Lett.* **156**, 129–32 (1997).
49. Touchon, M., Moura de Sousa, J. A. & Rocha, E. P. Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer. *Curr. Opin. Microbiol.* **38**, 66–73 (2017).
50. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: Building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
51. Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.-L. & Brüssow, H. Phage as agents of lateral gene transfer. *Curr. Opin. Microbiol.* **6**, 417–424 (2003).
52. Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801–812 (2007).
53. Takahashi, R. *et al.* High frequency of phage-infected bacterial cells in a rice field soil in Japan. *Soil Sci. Plant Nutr.* **57**, 35–39 (2011).
54. Bowatte, S., Newton, P. C. D., Takahashi, R. & Kimura, M. High frequency of virus-infected bacterial cells in a sheep grazed pasture soil in New Zealand. *Soil Biol. Biochem.* **42**, 708–712 (2010).
55. Munson-McGee, J. H. *et al.* A virus or more in (nearly) every cell: ubiquitous networks of virus–host interactions in extreme environments. *ISME J.* **12**, 1706–1714 (2018).
56. Danovaro, R. *et al.* Major viral impact on the functioning of benthic deep-sea ecosystems. *Nature* **454**, 1084–1087 (2008).
57. Breitbart, M. Marine Viruses: Truth or Dare. *Ann. Rev. Mar. Sci.* **4**, 425–448 (2012).
58. Kuzyakov, Y. & Mason-Jones, K. Viruses in soil: Nano-scale undead drivers of microbial life, biogeochemical turnover and ecosystem functions. *Soil Biol. Biochem.* **127**, 305–317 (2018).
59. Weinbauer, M. G. & Rassoulzadegan, F. Are viruses driving microbial diversification and diversity? *Environ. Microbiol.* **6**, 1–11 (2004).
60. Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **25**, 195–209 (2019).

61. Shreiner, A. B., Kao, J. Y. & Young, V. B. The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.* **31**, 69–75 (2015).
62. Edwards, R. A. *et al.* Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat. Microbiol.* 527796 (2019). doi:10.1038/s41564-019-0494-6
63. De Sordi, L., Lourenço, M. & Debarbieux, L. The Battle Within: Interactions of Bacteriophages and Bacteria in the Gastrointestinal Tract. *Cell Host Microbe* **25**, 210–218 (2019).
64. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
65. Manrique, P. *et al.* Healthy human gut phageome. *Proc. Natl. Acad. Sci.* **113**, 10400–10405 (2016).
66. Norman, J. M. *et al.* Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease. *Cell* **160**, 447–460 (2015).
67. Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* **26**, 764-778.e5 (2019).
68. Zuo, T. *et al.* Gut mucosal virome alterations in ulcerative colitis. *Gut* **68**, 1169–1179 (2019).
69. Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc. Natl. Acad. Sci.* **112**, 11941–11946 (2015).
70. Oh, J.-H. *et al.* Dietary Fructose and Microbiota-Derived Short-Chain Fatty Acids Promote Bacteriophage Production in the Gut Symbiont Lactobacillus reuteri. *Cell Host Microbe* **25**, 273-284.e6 (2019).
71. Filée, J., Tétart, F., Suttle, C. A. & Krisch, H. M. Marine T4-type bacteriophages, a ubiquitous component of the dark matter of the biosphere. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 12471–12476 (2005).
72. Walker, P. J. *et al.* Changes to virus taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2019). *Arch. Virol.* **164**, 2417–2429 (2019).
73. Krishnamurthy, S. R., Janowski, A. B., Zhao, G., Barouch, D. & Wang, D. Hyperexpansion of RNA Bacteriophage Diversity. *PLoS Biol.* **14**, e1002409 (2016).
74. Labonté, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* **7**, 2169–2177 (2013).
75. Creasy, A., Rosario, K., Leigh, B. A., Dishaw, L. J. & Breitbart, M. Unprecedented

diversity of ssDNA phages from the family Microviridae detected within the gut of a protochordate model organism (Ciona robusta). *Viruses* **10**, (2018).

76. Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118–122 (2018).
77. Roux, S. *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat. Microbiol.* **4**, 1895–1906 (2019).
78. Rappé, M. S. & Giovannoni, S. J. The Uncultured Microbial Majority. *Annu. Rev. Microbiol.* **57**, 369–394 (2003).
79. Brown, C. T. *et al.* Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**, 208–211 (2015).
80. Paez-Espino, D. *et al.* IMG/VR: A database of cultured and uncultured DNA viruses and retroviruses. *Nucleic Acids Res.* **45**, D457–D465 (2017).
81. Abedon, S. T. & Yin, J. Bacteriophage Plaques: Theory and Analysis. in *Bacteriophages: Methods and Protocols, Volume 1: Isolation, Characterization, and Interactions* **501**, 161–174 (2009).
82. Brum, J. R. *et al.* Illuminating structural proteins in viral "dark matter" with metaproteomics. *Proc. Natl. Acad. Sci.* **113**, 2436–2441 (2016).
83. Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* **5**, 1–11 (2014).
84. Reyes, A. *et al.* Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* **466**, 334–338 (2010).
85. Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653-664.e6 (2018).
86. Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat. Microbiol.* **3**, 38–46 (2018).
87. Shkoporov, A. N. *et al.* ΦCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat. Commun.* **9**, 4781 (2018).
88. Džunková, M. *et al.* Defining the human gut host–phage network through single-cell viral tagging. *Nat. Microbiol.* **4**, 2192–2203 (2019).
89. Rothenberg, E. *et al.* Single-virus tracking reveals a spatial receptor-dependent search mechanism. *Biophys. J.* **100**, 2875–2882 (2011).
90. Moldovan, R., Chapman-McQuiston, E. & Wu, X. L. On kinetics of phage adsorption. *Biophys. J.* **93**, 303–315 (2007).
91. Evilevitch, A., Lavelle, L., Knobler, C. M., Raspaud, E. & Gelbart, W. M. Osmotic

pressure inhibition of DNA ejection from phage. *Proc. Natl. Acad. Sci.* **100**, 9292–9295 (2003).

92. Edwards, R. A., McNair, K., Faust, K., Raes, J. & Dutilh, B. E. Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.* **40**, 258–272 (2016).
93. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).
94. Burstein, D. *et al.* Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat. Commun.* **7**, 1–8 (2016).
95. Manrique, P., Dills, M. & Young, M. J. The human gut phage community and its implications for health and disease. *Viruses* **9**, 9–11 (2017).
96. Zhao, G. *et al.* Intestinal virome changes precede autoimmunity in type I diabetes-susceptible children. *Proc. Natl. Acad. Sci.* **114**, E6166–E6175 (2017).
97. Nakatsu, G. *et al.* Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology* **155**, 529-541.e5 (2018).
98. Larsbrink, J. *et al.* A discrete genetic locus confers xyloglucan metabolism in select human gut Bacteroidetes. *Nature* **506**, 498–502 (2014).
99. Roche, D. B. *et al.* Classification of β-hairpin repeat proteins. *J. Struct. Biol.* **201**, 130–138 (2018).
100. Fraser, J. S., Maxwell, K. L. & Davidson, A. R. Immunoglobulin-like domains on bacteriophage: weapons of modest damage? *Curr. Opin. Microbiol.* **10**, 382–387 (2007).
101. Hayes, S. *et al.* Ubiquitous Carbohydrate Binding Modules Decorate 936 Lactococcal Siphophage Virions. *Viruses* **11**, 631 (2019).
102. Granell, M., Namura, M., Alvira, S., Kanamaru, S. & van Raaij, M. J. Crystal structure of the carboxy-terminal region of the bacteriophage T4 proximal long tail fiber protein Gp34. *Viruses* **9**, (2017).
103. Bartual, S. G. *et al.* Structure of the bacteriophage T4 long tail fiber receptor-binding tip. *Proc. Natl. Acad. Sci.* **107**, 20287–20292 (2010).
104. Spinelli, S. *et al.* Lactococcal bacteriophage p2 receptor-binding protein structure suggests a common ancestor gene with bacterial and mammalian viruses. *Nat. Struct. Mol. Biol.* **13**, 85–89 (2006).
105. Barbirz, S. *et al.* Crystal structure of Escherichia coli phage HK620 tailspike: Podoviral tailspike endoglycosidase modules are evolutionarily related. *Mol. Microbiol.* **69**, 303–316 (2008).

106. Steinbacher, S. *et al.* Phage P22 tailspike protein: Crystal structure of the head-binding domain at 2.3 Å, fully refined structure of the endorhamnosidase at 1.56 Å resolution, and the molecular basis of O-antigen recognition and cleavage. *J. Mol. Biol.* **267**, 865–880 (1997).
107. Steinbacher, S. *et al.* Crystal structure of P22 tailspike protein: Interdigitated subunits in a thermostable trimer. *Science (80-.).* **265**, 383–386 (1994).
108. Müller, J. J. *et al.* An Intersubunit Active Site between Supercoiled Parallel β Helices in the Trimeric Tailspike Endorhamnosidase of Shigella flexneri Phage Sf6. *Structure* **16**, 766–775 (2008).
109. Mitraki, A., Papanikolopoulou, K. & Van Raaij, M. J. Natural Triple β-Stranded Fibrous Folds. in *Advances in Protein Chemistry* **73**, 97–124 (2006).
110. Jernigan, K. K. & Bordenstein, S. R. Tandem-repeat protein domains across the tree of life. *PeerJ* **2015**, 1–16 (2015).
111. Verstrepen, K. J., Jansen, A., Lewitter, F. & Fink, G. R. Intragenic tandem repeats generate functional variability. *Nat. Genet.* **37**, 986–990 (2005).
112. Björklund, Å. K., Ekman, D. & Elofsson, A. Expansion of protein domain repeats. *PLoS Comput. Biol.* **2**, 0959–0970 (2006).
113. Wright, C. F., Teichmann, S. A., Clarke, J. & Dobson, C. M. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* **438**, 878–881 (2005).
114. Persi, E., Wolf, Y. I. & Koonin, E. V. Positive and strongly relaxed purifying selection drive the evolution of repeats in proteins. *Nat. Commun.* **7**, 1–11 (2016).
115. Finn, R. D. *et al.* Pfam: The protein families database. *Nucleic Acids Res.* **42**, 222–230 (2014).
116. Mello, L. V, Chen, X. & Rigden, D. J. Mining metagenomic data for novel domains: BACON, a new carbohydrate-binding module. *FEBS Lett.* **584**, 2421–2426 (2010).
117. Wheeler, T. J., Clements, J. & Finn, R. D. Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics* **15**, 1–9 (2014).
118. Rahmann, S., Schuster-Böckler, B. & Schultz, J. HMM logos for visualization of protein families. *BMC Bioinformatics* **5**, 7 (2004).
119. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425–430 (2016).
120. Jahn, M. T. *et al.* A Phage Protein Aids Bacterial Symbionts in Eukaryote Immune Evasion. *Cell Host Microbe* **26**, 542-550.e5 (2019).
121. Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate Representation

of Evolutionary and Functional Relationships between Phage Genomes. *Mol. Biol. Evol.* **25**, 762–777 (2008).

122. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).
123. Low, S. J., Džunková, M., Chaumeil, P.-A., Parks, D. H. & Hugenholtz, P. Evaluation of a concatenated protein phylogeny for classification of tailed double-stranded DNA viruses belonging to the order Caudovirales. *Nat. Microbiol.* **4**, 1306–1315 (2019).
124. Serwer, P. *et al.* Improved isolation of undersampled bacteriophages: Finding of distant terminase genes. *Virology* **329**, 412–424 (2004).
125. Toussaint, A. & Rice, P. A. Transposable phages, DNA reorganization and transfer. *Curr. Opin. Microbiol.* **38**, 88–94 (2017).
126. Fogg, P. C. M., Rigden, D. J., Saunders, J. R., McCarthy, A. J. & Allison, H. E. Characterization of the relationship between integrase, excisionase and antirepressor activities associated with a superinfecting Shiga toxin encoding bacteriophage. *Nucleic Acids Res.* **39**, 2116–2129 (2011).
127. Galiez, C., Siebert, M., Enault, F., Vincent, J. & Söding, J. WIsH: who is the host? Predicting prokaryotic hosts from metagenomic phage contigs. *Bioinformatics* **33**, 3113–3114 (2017).
128. de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J. & Dutilh, B. E. Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends Microbiol.* **27**, 51–63 (2019).
129. Garcia-Doval, C. & van Raaij, M. J. Structure of the receptor-binding carboxy-terminal domain of bacteriophage T7 tail fibers. *Proc. Natl. Acad. Sci.* **109**, 9390–9395 (2012).
130. Lasica, A. M., Ksiazek, M., Madej, M. & Potempa, J. The Type IX Secretion System (T9SS): Highlights and Recent Insights into Its Structure and Function. *Front. Cell. Infect. Microbiol.* **7**, (2017).
131. Fraser, J. S., Yu, Z., Maxwell, K. L. & Davidson, A. R. Ig-Like Domains on Bacteriophages: A Tale of Promiscuity and Deceit. *J. Mol. Biol.* **359**, 496–507 (2006).
132. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
133. Shiffman, M. E. *et al.* Gene and genome-centric analyses of koala and wombat fecal microbiomes point to metabolic specialization for Eucalyptus digestion. *PeerJ* **5**, e4075 (2017).

134. Roux, S. *et al.* Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **537**, 689–693 (2016).
135. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, (2010).
136. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, (2011).
137. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).
138. Enright, A. J. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
139. Charrad, M., Ghazzali, N., Boiteau, V. & Niknafs, A. NbClust : An R Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Softw.* **61**, 11744–11750 (2014).
140. Galili, T., O'Callaghan, A., Sidi, J. & Sievert, C. Heatmaply: An R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics* **34**, 1600–1602 (2018).
141. Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Sci.* **27**, 135–145 (2018).
142. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
143. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
144. Sullivan, M. J., Petty, N. K. & Beatson, S. A. Easyfig: A genome comparison visualizer. *Bioinformatics* **27**, 1009–1010 (2011).
145. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 1–9 (2009).
146. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
147. Zhou, X., Shen, X. X., Hittinger, C. T. & Rokas, A. Evaluating fast maximum likelihood-based phylogenetic programs using empirical phylogenomic data sets. *Mol. Biol. Evol.* **35**, 486–503 (2018).
148. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

149. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular biology and evolution. *Mol. Biol. Evol.* **35**, 518–522 (2018).
150. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* **47**, W256–W259 (2019).
151. Wattam, A. R. *et al.* PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.* **42**, 581–591 (2014).
152. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
153. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, 61–65 (2007).
154. Johnson, M. *et al.* NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, 5–9 (2008).
155. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
156. O'Sullivan, L., Buttimer, C., McAuliffe, O., Bolton, D. & Coffey, A. Bacteriophage-based tools: recent advances and novel applications. *F1000Research* **5**, 2782 (2016).
157. Zablocki, O., Adriaenssens, E. M. & Cowan, D. Diversity and Ecology of Viruses in Hyperarid Desert Soils. *Appl. Environ. Microbiol.* **82**, 770–777 (2016).
158. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science (80-.).* **348**, 1261498–1261498 (2015).
159. Weitz, J. S., Beckett, S. J., Brum, J. R., Cael, B. B. & Dushoff, J. Lysis, lysogeny and virus–microbe ratios. *Nature* **549**, E1–E3 (2017).
160. Knowles, B. *et al.* Lytic to temperate switching of viral communities. *Nature* **531**, 466–470 (2016).
161. Yu, P., Mathieu, J., Yang, Y. & Alvarez, P. J. J. Suppression of Enteric Bacteria by Bacteriophages: Importance of Phage Polyvalence in the Presence of Soil Bacteria. *Environ. Sci. Technol.* **51**, 5270–5278 (2017).
162. Modi, S. R., Lee, H. H., Spina, C. S. & Collins, J. J. Antibiotic treatment expands the resistance reservoir and ecological network of the phage metagenome. *Nature* **499**, 219–22 (2013).
163. Moebus, K. & Nattkemper, H. Bacteriophage sensitivity patterns among bacteria isolated

from marine waters. *Helgoländer Meeresuntersuchungen* **34**, 375–385 (1981).

164. Dekel-Bird, N. P., Sabehi, G., Mosevitzky, B. & Lindell, D. Host-dependent differences in abundance, composition and host range of cyanophages from the Red Sea. *Environ. Microbiol.* **17**, 1286–1299 (2015).
165. Siringan, P., Connerton, P. L., Cummings, N. J. & Connerton, I. F. Alternative bacteriophage life cycles: the carrier state of Campylobacter jejuni. *Open Biol.* **4**, 130200–130200 (2014).
166. Cenens, W., Makumi, A., Mebrhatu, M. T., Lavigne, R. & Aertsen, A. Phage–host interactions during pseudolysogeny. *Bacteriophage* **3**, e25029 (2013).
167. Peterson, A. Defining viral species: making taxonomy useful. *Virol. J.* **11**, 131 (2014).
168. Rosselló-Móra, R. & Amann, R. Past and future species definitions for Bacteria and Archaea. *Syst. Appl. Microbiol.* **38**, 209–216 (2015).
169. Serwer, P., Hayes, S. J., Thomas, J. A. & Hardies, S. C. Propagating the missing bacteriophages: a large bacteriophage in a new class. *Virol. J.* **4**, 21 (2007).
170. Meyer, J. R. *et al.* Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science (80-.).* **6056**, 1–8 (2016).
171. Schwarzer, D. *et al.* A Multivalent Adsorption Apparatus Explains the Broad Host Range of Phage phi92: a Comprehensive Genomic and Structural Analysis. *J. Virol.* **86**, 10384–10398 (2012).
172. Tu, J. *et al.* Dual host specificity of phage SP6 is facilitated by tailspike rotation. *Virology* **507**, 206–215 (2017).
173. Dowah, A. S. A. A. & Clokie, M. R. J. J. Review of the nature, diversity and structure of bacteriophage receptor binding proteins that target Gram-positive bacteria. *Biophys. Rev.* **1**, 1–8 (2018).
174. Silva, J. B. *et al.* Host receptors for bacteriophage adsorption. *FEMS Microbiol. Lett.* **363**, 1–11 (2016).
175. Takeuchi, I. *et al.* The presence of two receptor-binding proteins contributes to the wide host range of staphylococcal twort-like phages. *Appl. Environ. Microbiol.* **82**, 5763–5774 (2016).
176. Liu, M. Reverse Transcriptase-Mediated Tropism Switching in Bordetella Bacteriophage. *Science (80-.).* **295**, 2091–2094 (2002).
177. Perry, E. B., Barrick, J. E. & Bohannan, B. J. M. The Molecular and Genetic Basis of Repeatable Coevolution between Escherichia coli and Bacteriophage T3 in a Laboratory Microcosm. *PLoS One* **10**, e0130639 (2015).

178. Meyer, J. R. *et al.* Repeatability and Contingency in the Evolution of a Key Innovation in Phage Lambda. *Science (80-.).* **335**, 428–432 (2012).
179. Viana, D. *et al.* A single natural nucleotide mutation alters bacterial pathogen host tropism. *Nat. Genet.* **47**, 361–366 (2015).
180. Marston, M. F. *et al.* Rapid diversification of coevolving marine Synechococcus and a virus. *Proc. Natl. Acad. Sci.* **109**, 4544–4549 (2012).
181. Le, S. *et al.* Mapping the Tail Fiber as the Receptor Binding Protein Responsible for Differential Host Specificity of Pseudomonas aeruginosa Bacteriophages PaP1 and JG004. *PLoS One* **8**, 1–8 (2013).
182. Schwartz, D. A. & Lindell, D. Genetic hurdles limit the arms race between Prochlorococcus and the T7-like podoviruses infecting them. *ISME J.* **11**, 1836–1851 (2017).
183. Petrie, K. L. *et al.* Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. *Science (80-.).* **359**, 1542–1545 (2018).
184. Tétart, F., Repoila, F., Monod, C. & Krisch, H. M. Bacteriophage T4 Host Range is Expanded by Duplications of a Small Domain of the Tail Fiber Adhesin. *J. Mol. Biol.* **258**, 726–731 (1996).
185. Minot, S., Grunberg, S., Wu, G. D., Lewis, J. D. & Bushman, F. D. Hypervariable loci in the human gut virome. *Proc. Natl. Acad. Sci.* **109**, 3962–3966 (2012).
186. Paul, B. G. *et al.* Targeted diversity generation by intraterrestrial archaea and archaeal viruses. *Nat. Commun.* **6**, 6585 (2015).
187. Chow, L. T. & Bukhari, A. I. The invertible DNA segments of coliphages Mu and P1 are identical. *Virology* **74**, 242–248 (1976).
188. Golomidova, A. K. *et al.* Branched lateral tail fiber organization in T5-like bacteriophages DT57C and DT571/2 is revealed by genetic and functional analysis. *Viruses* **8**, 1–21 (2016).
189. Johnson, M. C. *et al.* Sinorhizobium meliloti Phage ΦM9 Defines a New Group of T4 Superfamily Phages with Unusual Genomic Features but a Common T=16 Capsid. *J. Virol.* **89**, 10945–58 (2015).
190. Tzipilevich, E., Habusha, M. & Ben-Yehuda, S. Acquisition of Phage Sensitivity by Bacteria through Exchange of Phage Receptors. *Cell* **168**, 186-199.e12 (2017).
191. Turnbull, L. *et al.* Explosive cell lysis as a mechanism for the biogenesis of bacterial membrane vesicles and biofilms. *Nat. Commun.* **7**, 11220 (2016).
192. Ando, H., Lemire, S., Pires, D. P. & Lu, T. K. Engineering Modular Viral Scaffolds for

Targeted Bacterial Population Editing. *Cell Syst.* **1**, 187–196 (2015).

193. Chen, M. *et al.* Alterations in gp37 Expand the Host Range of a T4-Like Phage. *Appl. Environ. Microbiol.* **83**, e01576-17 (2017).
194. Yosef, I., Goren, M. G., Globus, R., Molshanski-Mor, S. & Qimron, U. Extending the Host Range of Bacteriophage Particles for DNA Transduction. *Mol. Cell* **66**, 721-728.e3 (2017).
195. Samson, J. E., Magadán, A. H., Sabri, M. & Moineau, S. Revenge of the phages: defeating bacterial defences. *Nat. Rev. Microbiol.* **11**, 675–687 (2013).
196. Goldfarb, T. *et al.* BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.* **34**, 169–183 (2015).
197. Ofir, G. *et al.* DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.* **3**, 90–98 (2018).
198. Pawluk, A. *et al.* Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol.* **1**, 1–6 (2016).
199. Rifat, D., Wright, N. T., Varney, K. M., Weber, D. J. & Black, L. W. Restriction Endonuclease Inhibitor IPI* of Bacteriophage T4: A Novel Structure for a Dedicated Target. *J. Mol. Biol.* **375**, 720–734 (2008).
200. Chaikeeratisak, V. *et al.* Assembly of a nucleus-like structure during viral replication in bacteria. *Science (80-.).* **355**, 194–197 (2017).
201. Iyer, L. M., Burroughs, A. M., Anand, S., de Souza, R. F. & Aravind, L. Polyvalent proteins, a pervasive theme in the intergenomic biological conflicts of bacteriophages and conjugative elements. *J. Bacteriol.* **199**, 1–31 (2017).
202. Doron, S. *et al.* Transcriptome dynamics of a broad host-range cyanophage and its hosts. *ISME J.* **10**, 1437–1455 (2016).
203. Howard-Varona, C. *et al.* Regulation of infection efficiency in a globally abundant marine Bacteriodetes virus. *ISME J.* **11**, 284–295 (2017).
204. Howard-Varona, C. *et al.* Multiple mechanisms drive phage infection efficiency in nearly identical hosts. *ISME J.* **12**, 1605–1618 (2018).
205. Blasdel, B. G., Chevallereau, A., Monot, M., Lavigne, R. & Debarbieux, L. Comparative transcriptomics analyses reveal the conservation of an ancestral infectious strategy in two bacteriophage genera. *ISME J.* **11**, 1988–1996 (2017).
206. Williams, K. P. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Res.* **30**, 866–75 (2002).

207. Hammerl, J. A. *et al.* Analysis of the first temperate broad host range Brucellaphage (BiPBO1) isolated from B. inopinata. *Front. Microbiol.* **7**, 1–13 (2016).
208. Gilcrease, E. B. & Casjens, S. R. The genome sequence of Escherichia coli tailed phage D6 and the diversity of Enterobacteriales circular plasmid prophages. *Virology* **515**, 203–214 (2018).
209. Jäckel, C. *et al.* Prevalence, host range, and comparative genomic analysis of temperate Ochrobactrum phages. *Front. Microbiol.* **8**, 1–16 (2017).
210. Loessner, M. J. Bacteriophage endolysins - Current state of research and applications. *Curr. Opin. Microbiol.* **8**, 480–487 (2005).
211. Kong, M. & Ryu, S. Bacteriophage PBC1 and its endolysin as an antimicrobial agent against Bacillus cereus. *Appl. Environ. Microbiol.* **81**, 2274–2283 (2015).
212. Roces, C. *et al.* Reduced binding of the endolysin LYsTP712 to Lactococcus lactis ΔftsH contributes to phage resistance. *Front. Microbiol.* **7**, 1–10 (2016).
213. Ford, B. E. *et al.* Frequency and Fitness Consequences of Bacteriophage Φ6 Host Range Mutations. *PLoS One* **9**, e113078 (2014).
214. Keen, E. C. Tradeoffs in bacteriophage life histories. *Bacteriophage* **4**, e28365 (2014).
215. Avrani, S., Schwartz, D. A. & Lindell, D. Virus-host swinging party in the oceans. *Mob. Genet. Elements* **2**, 88–95 (2012).
216. Heineman, R. H., Springman, R. & Bull, J. J. Optimal Foraging by Bacteriophages through Host Avoidance. *Am. Nat.* **171**, E149–E157 (2008).
217. Benmayor, R., Hodgson, D. J., Perron, G. G. & Buckling, A. Host Mixing and Disease Emergence. *Curr. Biol.* **19**, 764–767 (2009).
218. Guyader, S. & Burch, C. L. Optimal Foraging Predicts the Ecology but Not the Evolution of Host Specialization in Bacteriophages. *PLoS One* **3**, e1946 (2008).
219. Bono, L. M., Gensel, C. L., Pfennig, D. W. & Burch, C. L. Competition and the origins of novelty: experimental evolution of niche-width expansion in a virus. *Biol. Lett.* **9**, 20120616–20120616 (2012).
220. Brum, J. R., Schenck, R. O. & Sullivan, M. B. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* **7**, 1738–1751 (2013).
221. Yoshida, M. *et al.* Quantitative viral community DNA analysis reveals the dominance of single-stranded DNA viruses in offshore upper bathyal sediment from Tohoku, Japan. *Front. Microbiol.* **9**, 1–10 (2018).
222. Hyman, P. & Abedon, S. T. Bacteriophage Host Range and Bacterial Resistance. in

Advances in Applied Microbiology **70**, 217–248 (Elsevier Inc., 2010).

223. Bielke, L., Higgins, S., Donoghue, A., Donoghue, D. & Hargis, B. M. Salmonella Host Range of Bacteriophages That Infect Multiple Genera. *Poult. Sci.* **86**, 2536–2540 (2007).
224. Jensen, E. C. *et al.* Prevalence of broad-host-range lytic bacteriophages of Sphaerotilus natans, Escherichia coli, and Pseudomonas aeruginosa. *Appl. Environ. Microbiol.* **64**, 575–580 (1998).
225. Yu, P., Mathieu, J., Li, M., Dai, Z. & Alvarez, P. J. J. J. Isolation of Polyvalent Bacteriophages by Sequential Multiple-Host Approaches. *Appl. Environ. Microbiol.* **82**, 808–815 (2016).
226. Tadmor, A. D., Ottesen, E. A., Leadbetter, J. R. & Phillips, R. Probing Individual Environmental Bacteria for Viruses by Using Microfluidic Digital PCR. *Science (80-.).* **333**, 58–62 (2011).
227. Allers, E. *et al.* Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environ. Microbiol.* **15**, 2306–2318 (2013).
228. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sci. Adv.* **3**, e1602105 (2017).
229. Cisek, A. A., Dąbrowska, I., Gregorczyk, K. P. & Wyżewski, Z. Phage Therapy in Bacterial Infections Treatment: One Hundred Years After the Discovery of Bacteriophages. *Curr. Microbiol.* **74**, 277–283 (2017).
230. Bordenstein, S. R. & Bordenstein, S. R. Novel eukaryotic association module in phage WO genomes from Wolbachia. *Nat. Commun.* **7**, 1–10 (2016).
231. Jover, L. F., Romberg, J. & Weitz, J. S. Inferring phage–bacteria infection networks from time-series data. *R. Soc. Open Sci.* **3**, 160654 (2016).
232. Sieber, M. & Gudelj, I. Do-or-die life cycles and diverse post-infection resistance mechanisms limit the evolution of parasite host ranges. *Ecol. Lett.* **17**, 491–498 (2014).
233. Cowley, L. A. *et al.* Analysis of whole genome sequencing for the Escherichia coli O157:H7 typing phages. *BMC Genomics* **16**, 271 (2015).
234. Endersen, L. *et al.* Phage therapy in the food industry. *Annu. Rev. Food Sci. Technol.* **5**, 327–49 (2014).
235. Schooley, R. T. *et al.* Development and use of personalized bacteriophage-based therapeutic cocktails to treat a patient with a disseminated resistant Acinetobacter baumannii infection. *Antimicrob. Agents Chemother.* **61**, e00954-17 (2017).

236. Soffer, N., Woolston, J., Li, M., Das, C. & Sulakvelidze, A. Bacteriophage preparation lytic for Shigella significantly reduces Shigella sonnei contamination in various foods. *PLoS One* **12**, 1–11 (2017).
237. Drulis-Kawa, Z., Majkowska-Skrobek, G., Maciejewska, B., Delattre, A.-S. & Lavigne, R. Learning from Bacteriophages - Advantages and Limitations of Phage and Phage-Encoded Protein Applications. *Curr. Protein Pept. Sci.* **13**, 699–722 (2012).
238. Nobrega, F. L., Costa, A. R., Kluskens, L. D. & Azeredo, J. Revisiting phage therapy: New applications for old resources. *Trends Microbiol.* **23**, 185–191 (2015).
239. Chan, B. K., Abedon, S. T. & Loc-Carrillo, C. Phage cocktails and the future of phage therapy. *Future Microbiol.* **8**, 769–783 (2013).
240. Edgar, R., Friedman, N., Shahar, M. M. & Qimron, U. Reversing bacterial resistance to antibiotics by phage-mediated delivery of dominant sensitive genes. *Appl. Environ. Microbiol.* **78**, 744–751 (2012).
241. Yosef, I., Manor, M., Kiro, R. & Qimron, U. Temperate and lytic bacteriophages programmed to sensitize and kill antibiotic-resistant bacteria. *Proc. Natl. Acad. Sci.* **112**, 7267–7272 (2015).
242. Park, J. Y. *et al.* Genetic engineering of a temperate phage-based delivery system for CRISPR/Cas9 antimicrobials against Staphylococcus aureus. *Sci. Rep.* **7**, 1–13 (2017).
243. Korytowski, D. A. & Smith, H. L. How nested and monogamous infection networks in host-phage communities come to be. *Theor. Ecol.* **8**, 111–120 (2014).
244. Flores, C. O., Meyer, J. R., Valverde, S., Farr, L. & Weitz, J. S. Statistical structure of host-phage interactions. *Proc. Natl. Acad. Sci.* **108**, E288–E297 (2011).
245. Gomez, P. & Buckling, A. Bacteria-Phage Antagonistic Coevolution in Soil. *Science (80-.).* **332**, 106–109 (2011).
246. Hall, A. R., Scanlan, P. D., Morgan, A. D. & Buckling, A. Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecol. Lett.* **14**, 635–642 (2011).
247. Weitz, J. S. *et al.* Phage–bacteria infection networks. *Trends Microbiol.* **21**, 82–91 (2013).
248. Flores, C. O., Valverde, S. & Weitz, J. S. Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *ISME J.* **7**, 520–532 (2013).
249. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* **4**, e08490 (2015).
250. Scanlan, P. D., Hall, A. R., Burlinson, P., Preston, G. & Buckling, A. No effect of host-parasite co-evolution on host range expansion. *J. Evol. Biol.* **26**, 205–209 (2013).

251. Koskella, B. & Parr, N. The evolution of bacterial resistance against bacteriophages in the horse chestnut phyllosphere is general across both space and time. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140297 (2015).
252. Vos, M., Birkett, P. J., Birch, E., Griffiths, R. I. & Buckling, A. Local Adaptation of Bacteriophages to Their Bacterial Hosts in Soil. *Science (80-.).* **325**, 833–833 (2009).
253. Hanson, C. A., Marston, M. F. & Martiny, J. B. H. Biogeographic variation in host range phenotypes and taxonomic composition of marine cyanophage isolates. *Front. Microbiol.* **7**, 1–14 (2016).
254. Gurney, J. *et al.* Network structure and local adaptation in co-evolving bacteria–phage interactions. *Mol. Ecol.* **26**, 1764–1777 (2017).
255. Karimi, M. *et al.* Bacteriophages and phage-inspired nanocarriers for targeted delivery of therapeutic cargos. *Adv. Drug Deliv. Rev.* **106**, 45–62 (2016).
256. Chen, J. & Novick, R. P. Phage-Mediated Intergeneric Transfer of Toxin Genes. *Science (80-.).* **323**, 139–141 (2009).
257. Marcó, M. B., Moineau, S. & Quiberoni, A. Bacteriophages and dairy fermentations. *Bacteriophage* **2**, 149–158 (2012).
258. Corinaldesi, C., Dell'Anno, A. & Danovaro, R. Viral infections stimulate the metabolism and shape prokaryotic assemblages in submarine mud volcanoes. *ISME J.* **6**, 1250–1259 (2012).
259. Sharon, I. *et al.* Photosystem I gene cassettes are present in marine virus genomes. *Nature* **461**, 258–262 (2009).
260. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free d2∗ oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**, 39–53 (2017).
261. Liu, D., Ma, Y., Jiang, X. & He, T. Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion. *BMC Bioinformatics* **20**, 1–10 (2019).
262. Villarroel, J. *et al.* HostPhinder: A phage host prediction tool. *Viruses* **8**, 1–22 (2016).
263. Mihara, T. *et al.* Linking virus genomes with host taxonomy. *Viruses* **8**, 10–15 (2016).
264. Zhang, M. *et al.* Prediction of virus-host infectious association by supervised learning methods. *BMC Bioinformatics* **18**, (2017).
265. Deng, L. *et al.* Contrasting Life Strategies of Viruses that Infect Photo- and Heterotrophic Bacteria, as Revealed by Viral Tagging. *MBio* **3**, e00373-12-e00373-12 (2012).
266. Wang, J., Hofnung, M. & Charbit, A. The C-terminal portion of the tail fiber protein of

bacteriophage lambda is responsible for binding to LamB, its receptor at the surface of Escherichia coli K-12. *J. Bacteriol.* **182**, 508–512 (2000).

267. Andres, D. *et al.* Tailspike interactions with lipopolysaccharide effect DNA ejection from phage P22 particles in vitro. *J. Biol. Chem.* **285**, 36768–36775 (2010).
268. Randall-Hazelbauer, L. & Shwartz, M. Isolation of the bacteriophage lambda receptor from E. coli. *J. Bacteriol.* **116**, 1436–1446 (1973).
269. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 2006.0008 (2006).
270. Rice, L. B. Federal Funding for the Study of Antimicrobial Resistance in Nosocomial Pathogens: No ESKAPE. *J. Infect. Dis.* **197**, 1079–1081 (2008).
271. Pinard, R. *et al.* Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* **7**, (2006).
272. Poole, R. K. The isolation of membranes from bacteria. *Methods Mol. Biol.* **19**, 109–22 (1993).
273. Kim, K. H. & Bae, J. W. Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl. Environ. Microbiol.* **77**, 7663–7668 (2011).
274. Yilmaz, S., Allgaier, M. & Hugenholtz, P. Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat. Methods* **7**, 943–944 (2010).
275. Probst, A. J., Weinmaier, T., DeSantis, T. Z., Santo Domingo, J. W. & Ashbolt, N. New perspectives on microbial community distortion after whole-genome amplification. *PLoS One* **10**, 1–16 (2015).
276. Stopar, D., Spruijt, R. B., Wolfs, C. J. A. M. & Hemminga, M. A. Protein-lipid interactions of bacteriophage M13 major coat protein. *Biochim. Biophys. Acta - Biomembr.* **1611**, 5–15 (2003).
277. Wu, L. *et al.* Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat. Microbiol.* (2019). doi:10.1038/s41564-019-0426-5
278. Petrovich, M. L. *et al.* Viral composition and context in metagenomes from biofilm and suspended growth municipal wastewater treatment plants. *Microb. Biotechnol.* **12**, 1324–1336 (2019).
279. Agarwala, R. *et al.* Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **45**, D12–D17 (2017).
280. Nolan, J. M., Petrov, V., Bertrand, C., Krisch, H. M. & Karam, J. D. Genetic diversity among five T4-like bacteriophages. *Virol. J.* **3**, 1–15 (2006).

281. Wang, J.-B., Lin, N.-T., Tseng, Y.-H. & Weng, S.-F. Genomic Characterization of the Novel Aeromonas hydrophila Phage Ahp1 Suggests the Derivation of a New Subgroup from phiKMV-Like Family. *PLoS One* **11**, e0162060 (2016).
282. Comeau, A. M. *et al.* Phage morphology recapitulates phylogeny: The comparative genomics of a new group of myoviruses. *PLoS One* **7**, 1–11 (2012).
283. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327 (2010).
284. Hampton, H. G., Watson, B. N. J. & Fineran, P. C. The arms race between bacteria and their phage foes. *Nature* **577**, 327–336 (2020).
285. Bourkal'tseva, M. V. *et al.* Bacteriophage phi297, a new species of Pseudomonas aeruginosa temperate phages with a mosaic genome: Potential use in phage therapy. *Russ. J. Genet.* **47**, 794–798 (2011).
286. Lavigne, R., Seto, D., Mahadevan, P., Ackermann, H. W. & Kropinski, A. M. Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. *Res. Microbiol.* **159**, 406–414 (2008).
287. Hatfull, G. F. & Hendrix, R. W. Bacteriophages and their genomes. *Curr. Opin. Virol.* **1**, 298–303 (2011).
288. Garcia, E. *et al.* The genome sequence of Yersinia pestis bacteriophage φA1122 reveals an intimate history with the coliphage T3 and T7 genomes. *J. Bacteriol.* **185**, 5248–5262 (2003).
289. Yuan, Y. & Gao, M. Jumbo bacteriophages: An overview. *Front. Microbiol.* **8**, 1–9 (2017).
290. Yabuuchi, E. *et al.* Transfer of Two Burkholderia and An Alcaligenes Species to Ralstonia Gen. Nov. *Microbiol. Immunol.* **39**, 897–904 (1995).
291. Zdorovenko, E. L., Vinogradov, E., Wydra, K., Lindner, B. & Knirel, Y. A. Structure of the Oligosaccharide Chain of the SR-Type Lipopolysaccharide of Ralstonia solanacearum Toudk-2. *Biomacromolecules* **9**, 2215–2220 (2008).
292. Mendoza, S. D. *et al.* A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* **577**, (2019).
293. Arkhipova, K. *et al.* Temporal dynamics of uncultured viruses: A new dimension in viral diversity. *ISME J.* **12**, 199–211 (2018).
294. de Jonge, P. A., von Meijenfeldt, F. A. B., van Rooijen, L. E., Brouns, S. J. J. & Dutilh, B. E. Evolution of BACON Domain Tandem Repeats in crAssphage and Novel Gut Bacteriophage Lineages. *Viruses* **11**, 1085 (2019).

295. Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nat. Microbiol.* **2**, 1–9 (2017).
296. Metsky, H. C. *et al.* Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nat. Biotechnol.* **37**, 160–168 (2019).
297. Gómez-Rubio, V. ggplot2 - Elegant Graphics for Data Analysis (2nd Edition) . *J. Stat. Softw.* **77**, 3–5 (2017).
298. Allen-Vercoe, E., Dibb-Fuller, M., Thorns, C. J. & Woodward, M. J. SEF17 fimbriae are essential for the convoluted colonial morphology of Salmonella enteritidis. *FEMS Microbiol. Lett.* **153**, 33–42 (1997).
299. Kutter, E. Phage host range and efficiency of plating. *Methods Mol. Biol.* **501**, 141–9 (2009).
300. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
301. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
302. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
303. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
304. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **2015**, 1–15 (2015).
305. Von Meijenfeldt, F. A. B., Arkhipova, K., Cambuy, D. D., Coutinho, F. H. & Dutilh, B. E. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol.* **20**, 1–14 (2019).
306. Roux, S., Emerson, J. B., Eloe-Fadrosh, E. A. & Sullivan, M. B. Benchmarking viromics: An in silico evaluation of metagenome-enabled estimates of viral community composition and diversity. *PeerJ* **2017**, 1–26 (2017).
307. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123.e14 (2019).
308. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–504 (2003).
309. Grazziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic Virus Orthologous Groups (pVOGs): A resource for comparative genomics and protein family annotation.

Nucleic Acids Res. **45**, D491–D498 (2017).

310. Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. M.-L., Brussow, H. & Brüssow, H. Phage-Host Interaction: an Ecological Perspective. *J. Bacteriol.* **186**, 3677–3686 (2004).
311. Clokie, M. R. J., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31–45 (2011).
312. Koskella, B. New approaches to characterizing bacteria–phage interactions in microbial communities and microbiomes. *Environ. Microbiol. Rep.* **11**, 15–16 (2019).
313. São-José, C. *et al.* The ectodomain of the viral receptor YueB forms a fiber that triggers ejection of bacteriophage SPP1 DNA. *J. Biol. Chem.* **281**, 11464–11470 (2006).
314. Andres, D., Baxa, U., Hanke, C., Seckler, R. & Barbirz, S. Carbohydrate binding of Salmonella phage P22 tailspike protein and its role during host cell infection. *Biochem. Soc. Trans.* **38**, 1386–1389 (2010).
315. Broeker, N. K. *et al.* In vitro studies of lipopolysaccharide-mediated DNA release of podovirus HK620. *Viruses* **10**, (2018).
316. Wang, C., Tu, J., Liu, J. & Molineux, I. J. Structural dynamics of bacteriophage P22 infection initiation revealed by cryo-electron tomography. *Nat. Microbiol.* **4**, 1049–1056 (2019).
317. Rakhuba, D. V., Kolomiets, E. I., Szwajcer Dey, E. & Novik, G. I. Bacteriophage receptors, mechanisms of phage adsorption and penetration into host cell. *Polish J. Microbiol.* **59**, 145–155 (2010).
318. Lee, Y.-C. *et al.* Impact of Detergents on Membrane Protein Complex Isolation. *J. Proteome Res.* **17**, 348–358 (2018).
319. Gulati, S. *et al.* Detergent-free purification of ABC (ATP-binding-cassette) transporters. *Biochem. J.* **461**, 269–278 (2014).
320. Denisov, I. G. & Sligar, S. G. Nanodiscs for structural and functional studies of membrane proteins. *Nat. Struct. Mol. Biol.* **23**, 481–486 (2016).
321. Lee, S. C. *et al.* A method for detergent-free isolation of membrane proteins in their local lipid environment. *Nat. Protoc.* **11**, 1149–1162 (2016).
322. Dörr, J. M. *et al.* The styrene–maleic acid copolymer: a versatile tool in membrane research. *Eur. Biophys. J.* **45**, 3–21 (2016).
323. Jamshad, M. *et al.* Structural analysis of a nanoparticle containing a lipid bilayer used for detergent-free extraction of membrane proteins. *Nano Res.* **8**, 774–789 (2015).
324. Broecker, J., Eger, B. T. & Ernst, O. P. Crystallogenesis of Membrane Proteins Mediated by Polymer-Bounded Lipid Nanodiscs. *Structure* **25**, 384–392 (2017).

325. Bersch, B., Dörr, J. M., Hessel, A., Killian, J. A. & Schanda, P. Proton-Detected Solid-State NMR Spectroscopy of a Zinc Diffusion Facilitator Protein in Native Nanodiscs. *Angew. Chemie - Int. Ed.* **56**, 2508–2512 (2017).
326. Knowles, T. J. *et al.* Membrane proteins solubilized intact in lipid containing nanoparticles bounded by styrene maleic acid copolymer. *J. Am. Chem. Soc.* **131**, 7484–7485 (2009).
327. Sutcliffe, I. C. A phylum level perspective on bacterial cell envelope architecture. *Trends Microbiol.* **18**, 464–470 (2010).
328. Craig, A. F. *et al.* Tuning the size of styrene-maleic acid copolymer-lipid nanoparticles (SMALPs) using RAFT polymerization for biophysical studies. *Biochim. Biophys. Acta - Biomembr.* **1858**, 2931–2939 (2016).
329. Cuevas Arenas, R., Klingler, J., Vargas, C. & Keller, S. Influence of lipid bilayer properties on nanodisc formation mediated by styrene/maleic acid copolymers. *Nanoscale* **8**, 15016–15026 (2016).
330. Hall, S. C. L. *et al.* Influence of Poly(styrene- co -maleic acid) Copolymer Structure on the Properties and Self-Assembly of SMALP Nanodiscs. *Biomacromolecules* **19**, 761–772 (2018).
331. Gidden, J., Denson, J., Liyanage, R., Ivey, D. M. & Lay, J. O. Lipid compositions in Escherichia coli and Bacillus subtilis during growth as determined by MALDI-TOF and TOF/TOF mass spectrometry. *Int. J. Mass Spectrom.* **283**, 178–184 (2009).
332. Thwaites, J. J. & Surana, U. C. Mechanical properties of Bacillus subtilis cell walls: Effects of removing residual culture medium. *J. Bacteriol.* **173**, 197–203 (1991).
333. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb. Perspect. Biol.* **2**, a000414 (2010).
334. Washizaki, A., Yonesaki, T. & Otsuka, Y. Characterization of the interactions between Escherichia coli receptors, LPS and OmpC, and bacteriophage T4 long tail fibers. *Microbiologyopen* **5**, 1003–1015 (2016).
335. Bjornsti, M. A., Reilly, B. E. & Anderson, D. L. Morphogenesis of bacteriophage phi 29 of Bacillus subtilis: DNA-gp3 intermediate in in vivo and in vitro assembly. *J. Virol.* **41**, 508–17 (1982).
336. Shao, Y. & Wang, I. N. Effect of late promoter activity on bacteriophage λ fitness. *Genetics* **181**, 1467–1475 (2009).
337. Demerec, M. & Fano, U. Bacteriophage-Resistant Mutants in Escherichia Coli. *Genetics* **30**, 119–36 (1945).

338. Henry, M. *et al.* Development of a high throughput assay for indirectly measuring phage growth using the OmniLog(TM) system. *Bacteriophage* **2**, 159–167 (2012).
339. Fischer, W. Lipoteichoic acid and lipids in the membrane of Staphylococcus aureus. *Med. Microbiol. Immunol.* **183**, 61–76 (1994).
340. Chatterjee, S. & Rothenberg, E. Interaction of bacteriophage λ with Its E. coli receptor, LamB. *Viruses* **4**, 3162–3178 (2012).
341. Leiman, P. G. *et al.* Morphogenesis of the T4 tail and tail fibers. *Virol. J.* **7**, 355 (2010).
342. Yu, F. & Mizushima, S. Roles of lipopolysaccharide and outer membrane protein ompc of Escherichia coli K-12 in the receptor function for bacteriophage T4. *J. Bacteriol.* **151**, 718–722 (1982).
343. Islam, M. Z. *et al.* Molecular Anatomy of the Receptor Binding Module of a Bacteriophage Long Tail Fiber. *PLoS Pathog.* **15**, 1–21 (2019).
344. Lindhoud, S., Carvalho, V., Pronk, J. W. & Aubin-Tam, M. E. SMA-SH: Modified Styrene-Maleic Acid Copolymer for Functionalization of Lipid Nanodiscs. *Biomacromolecules* **17**, 1516–1522 (2016).
345. Postis, V. *et al.* The use of SMALPs as a novel membrane protein scaffold for structure study by negative stain electron microscopy. *Biochim. Biophys. Acta - Biomembr.* **1848**, 496–501 (2014).
346. Eriksson, M., Härdelin, M., Larsson, A., Bergenholtz, J. & Åkerman, B. Binding of intercalating and groove-binding cyanine dyes to bacteriophage T5. *J. Phys. Chem. B* **111**, 1139–1148 (2007).
347. Ravula, T. *et al.* Effect of polymer charge on functional reconstitution of membrane proteins in polymer nanodiscs. *Chem. Commun.* **54**, 9615–9618 (2018).
348. Rountree, P. M. The Role of Divalent Cations in the Multiplication of Staphylococcal Bacteriophages. *J. Gen. Microbiol.* **12**, 275–287 (1955).
349. Tolmach, L. J. & Puck, T. T. The Mechanism of Virus Attachment to Host Cells. *J. Am. Chem. Soc.* **74**, 5551–5553 (1952).
350. Abu Al-Soud, W. & Rådström, P. Capacity of nine thermostable DNA polymerases to mediate DNA amplification in the presence of PCR-inhibiting samples. *Appl. Environ. Microbiol.* **64**, 3748–3753 (1998).
351. Schrader, C., Schielke, A., Ellerbroek, L. & Johne, R. PCR inhibitors - occurrence, properties and removal. *J. Appl. Microbiol.* **113**, 1014–1026 (2012).
352. Rossen, L., Nørskov, P., Holmstrøm, K. & Rasmussen, O. F. Inhibition of PCR by components of food samples, microbial diagnostic assays and DNA-extraction solutions.

Int. J. Food Microbiol. **17**, 37–45 (1992).

353. Morfin, I. *et al.* Adsorption of divalent cations on DNA. *Biophys. J.* **87**, 2897–904 (2004).
354. Hall, S. C. L. *et al.* An acid-compatible co-polymer for the solubilization of membranes and proteins into lipid bilayer-containing nanoparticles. *Nanoscale* **10**, 10609–10619 (2018).
355. Ravula, T., Hardin, N. Z., Ramadugu, S. K., Cox, S. J. & Ramamoorthy, A. Formation of pH-Resistant Monodispersed Polymer–Lipid Nanodiscs. *Angew. Chemie - Int. Ed.* **57**, 1342–1345 (2018).
356. Taylor, N. M. I. *et al.* Structure of the T4 baseplate and its function in triggering sheath contraction. *Nature* **533**, 346–52 (2016).
357. Legrand, P. *et al.* The Atomic Structure of the Phage Tuc2009 Baseplate Tripod Suggests that Host Recognition Involves Two Different Carbohydrate Binding Modules. *MBio* **7**, e01781-15 (2016).
358. Thonghin, N., Kargas, V., Clews, J. & Ford, R. C. Cryo-electron microscopy of membrane proteins. *Methods* **147**, 176–186 (2018).
359. Weber, M., Wojtynek, M. & Medalia, O. Cellular and Structural Studies of Eukaryotic Cells by Cryo-Electron Tomography. *Cells* **8**, 57 (2019).
360. Fortier, L.-C. & Moineau, S. Phage Production and Maintenance of Stocks, Including Expected Stock Lifetimes. in *Methods in Gut Microbial Ecology for Ruminants* (eds. Clokie, M. R. J. & Kropinski, A. M.) **501**, 203–219 (Humana Press, 2009).
361. Lawrence, J. E. & Steward, G. F. Purification of viruses by centrifugation. *Man. Aquat. Viral Ecol. ASLO* **2**, 166–181 (2010).
362. Angelisová, P. *et al.* The use of styrene-maleic acid copolymer (SMA) for studies on T cell membrane rafts. *Biochim. Biophys. Acta - Biomembr.* **1861**, 130–141 (2019).
363. Kopf, A. H. *et al.* Factors influencing the solubilization of membrane proteins from Escherichia coli membranes by styrene–maleic acid copolymers. *Biochim. Biophys. Acta - Biomembr.* **1862**, 183125 (2020).
364. Rohwer, F. & Segall, A. M. In retrospect: A century of phage lessons. *Nature* **528**, 46–48 (2015).
365. Kurzgesagt. The Deadliest Being on Planet Earth – The Bacteriophage. Available at: https://www.youtube.com/watch?v=YI3tsmFsrOg&t=35s.
366. Yong, E. A Dying Teenager's Recovery Started in the Dirt. *the Atlantic* (2019). Available at: https://www.theatlantic.com/science/archive/2019/05/engineered-viruses-

cured-a-dying-girls-infections/589075/.

367. Corbyn, Z. Steffanie Strathdee: ‘Phages have evolved to become perfect predators of bacteria’. *the Observer* (2019). Available at: https://www.theguardian.com/science/2019/jun/15/steffanie-strathdee-phage-therapy-interview-perfect-predator.

368. Kraemer, J. A. *et al.* A phage tubulin assembles dynamic filaments by an atypical mechanism to center viral DNA within the host cell. *Cell* **149**, 1488–1499 (2012).

369. Silpe, J. E. & Bassler, B. L. A Host-Produced Quorum-Sensing Autoinducer Controls a Phage Lysis-Lysogeny Decision. *Cell* 1–13 (2018). doi:10.1016/j.cell.2018.10.059

370. Hoyland-Kroghsbo, N. M., Maerkedahl, R. B. & Svenningsen, S. Lo. A Quorum-Sensing-Induced Bacteriophage Defense Mechanism. *MBio* **4**, e00362-12-e00362-12 (2013).

371. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489–91 (2013).

372. Barr, J. J. *et al.* Subdiffusive motion of bacteriophage in mucosal surfaces increases the frequency of bacterial encounters. *Proc. Natl. Acad. Sci.* **112**, 13675–13680 (2015).

Samenvatting

Bacteriofagen (of fagen) zijn virussen die bacteriën infecteren. Er zijn tien keer meer fagen dan cellulaire organismen, en fagen zijn een onderdeel van elk ecosysteem op aarde. Hierdoor beïnvloeden ze alles: van de menselijke gezondheid tot en met de nutriëntcycli die aan de basis staan van al het leven op aarde. Ondanks hun klaarblijkelijke belang binnen de biologie, kwam inzicht in de effecten die fagen op allerlei ecosystemen hebben pas met de opkomst van metagenomica in de afgelopen twintig jaar. In metagenomica wordt het genetisch materiaal binnen een omgeving bemonsterd en onthuld. Dit laat het toe om fagen te bestuderen zonder de arbeidsintensieve en soms bijna onmogelijke taak om fagen te isoleren in een laboratorium.

Hoewel metagenomica heeft geresulteerd in een wedergeboorte van de faagbiologie, zijn er nog grote uitdagingen te slechten. Een noemenswaardige uitdaging is het om met voldoende zekerheid voor elke faag te kunnen achterhalen welke bacteriële soort ze infecteren, vooral als het gaat om bacteriën die moeilijk zijn om te isoleren en te cultiveren. Daarnaast zijn er ook fagen die door hun groei-eigenschappen niet kunnen worden geïsoleerd met de huidige beschikbare methoden. Het onderzoek in dit proefschrift is daarom gericht op het ontdekken en begrijpen van interacties tussen bacteriofagen en bacteriën.

Om de interacties tussen fagen en bacteriën beter te begrijpen, is de studie van de moleculaire mechanismen waarmee fagen aan hun gastheercellen binden zeer belangrijk. In **hoofdstuk 2** wordt een studie *Bacteroides*-gerelateerde koolhydraat-bindende meestal N-terminale (BACON) domeinen die binnen de genomen van crAssachtige fagen voorkomen beschreven. De crAssachtige faagfamilie is één van de meest voorkomende binnen de darmen van mensen wereldwijd, maar welk effect ze hebben op de menselijk darmflora is onbekend. De crAssachtige faag die als eerst ontdekt werd (crAssfaag) heeft een achtereenvolgende herhaling van BACON-domeinen waarover eerder de hypothese geopperd is dat deze een rol spelen in het binden van koolhydraten, bijvoorbeeld op het oppervlak van gastheercellen. Door deze domeinherhaling te bestuderen kregen we inzage in de evolutie van BACON-domeinen in crAssachtige fagen en hebben we meerdere nieuwe groepen fagen ontdekt.

Naast onze studie van BACON-domeinen in crAssachtige fagen bestudeerden we ook de moleculaire mechanismen waarmee fagen aan gastheercellen binden, zoals te lezen in **hoofdstuk 3**. In dit hoofdstuk wordt een synthese gepresenteerd van de huidige kennis over de verschillende aspecten die meespelen in de reikwijdte van gastheersoorten die fagen kunnen infecteren. Daarnaast bediscussiëren we de mechanismen die zijn geëvolueerd in fagen met een

brede reikwijdte van gastheersoorten. We beschrijven ook een nieuw model om de verschillen tussen gastheerwisselingen en daadwerkelijke brede reikwijdte van gastheersoorten te onderstrepen. Ten slotte laten we zien ook hoe huidige technieken de isolatie van fagen met een nauwe reikwijdte van gastheren bevooroordeeld.

De daaropvolgende twee hoofstukken richten zich op de ontwikkeling van nieuwe methodologieën waarmee de interacties tussen fagen en bacteriën kunnen worden bestudeerd. Adsorbatiesequencen (AdsorpSeq) wordt geïntroduceerd in **Hoofdstuk 4**. In AdsorpSeq wordt de interactie tussen fagen en de moleculen op het oppervlak van gastheercellen geïsoleerd. De methode bestaat uit vijf stappen. Eerst worden fagen en geïsoleerde bacteriële cel-enveloppen gemengd. Een incubatie laat fagen aan receptormoleculen op de bacteriële cel-enveloppen binden. In de derde stap worden de cel-enveloppen, en de daaraan gebonden fagen, gescheiden van ongebonden fagen. Dan wordt het DNA van de gebonden fagen geïsoleerd. De laatste stap bestaat ten slotte uit het sequencen en analyseren van dit DNA. Met deze vijf stappen is AdsorpSeq een zeer snelle en makkelijk uitvoerbare methode om de interacties tussen fagen en hun gastheercellen te bestuderen. Nadat we met enkele modelfagen laten zien dat AdsorpSeq werkt, gebruikten we de methode om interacties tussen fagen en bacteriën te bepalen in een monster uit een ziekenhuisafvalwaterzuiveringsinstallatie.

In **hoofdstuk 5** wordt, tot slot, de bruikbaarheid van styreen-maleïnezuur lipidedeeltjes (SMALPs) als potentieel nieuw gereedschap voor studies van interacties tussen fagen en bacteriën verkend. SMALPs zijn recent ontwikkeld als methode om bacteriële membraaneiwitten te isoleren zonder hun natuurlijke omgeving te verstoren. Dit is van belang omdat dergelijke membraaneiwitten meestal niet intact blijven buiten hun lipidemembranen en verder omdat ze een voorname categorie van faagreceptoren vormen. In het hier gepresenteerde project laten we zien dat verschillende modelfagen binden aan SMALPs die vervaardigd zijn van membranen van hun gastheren. Deze interacties resulteren in de ejectie van faaggenomen, net zoals gebeurt in het geval van een infectie van een intacte cel. De binding van fagen aan SMALPs vindt alleen plaats wanneer de SMALPs gemaakt zijn uit het membraan van de oorspronkelijke bacteriële gastheersoort, en dan alleen als specifieke receptormoleculen aanwezig zijn. We concluderen uiteindelijk dat SMALPs, rekening houdend met enkele zaken, potentieel toepasbaar zijn in de studie van interacties tussen fagen en bacteriën.

Samengevat levert dit proefschrift inzage in de vele methoden die in fagen geëvolueerd zijn om aan bacteriën te binden, met een gedetailleerde blik op de staarteiwitten van een wijdverbreide darmfaag. Daarnaast presenteert het twee nieuw methoden waarmee faag-bacterie interacties kunnen worden bestudeerd.

Summary

Bacteriophages (or phages) are viruses that infect bacteria. Phages outnumber cellular life tenfold and are present in every ecosystem on Earth. Thereby their impact extends from human health to the nutrient cycles that are the basis of life on Earth. Despite their evident importance in biology, much of the extent to which phages impact various ecosystems is only being uncovered since the advent of metagenomic sequencing in the early 2000s. In metagenomic sequencing, the genetic makeup from an environment is sampled, which enables studies of phages without the laborious and sometimes near impossible task of physically isolating them in a laboratory.

While metagenomic sequencing has resulted in a renaissance of phage biology, sizeable challenges remain. One notable challenge is to reliably link phages to the bacterial host species that they infect and on which they are completely dependent for their existence. This is especially true for phages infecting bacteria that are difficult or impossible to isolate and cultivate. In addition, many phages cannot be isolated with current methods due to their infection characteristics. This thesis therefore focuses on uncovering and understanding interactions between bacteriophages and bacteria.

One method to further understand phage-host interactions is to study the molecular mechanisms by which they interact with their hosts. **Chapter 2** details the study of putative host-binding *Bacteroides*-associated carbohydrate-binding often N-terminal (BACON) domains in crAss-like phages. The crAss-like phage lineage is among the most widespread phage lineages in the human gut, but its impact on the gut microbiome remains mysterious. The prototypical crAss-like phage (crAssphage) contains a BACON domain tandem repeat which was previously hypothesised to be involved in binding carbohydrates, for instance on the host cell surface. We studied these domain tandem repeats to further understanding of BACON domain evolution in crAss-like phages. Besides providing insight into crAss-like phage evolution, this study uncovered several novel phage lineages.

In addition to a study of crAss-like phage BACON domains, we further studied the molecular mechanisms that are used by phages to bind to their hosts, as reported in **chapter 3**. This chapter provides a synthesis of current knowledge on various aspects that govern phage host range and discusses the diversity of molecular mechanisms that phages possess to allow a broad host range. We additionally posit a novel framework to address the difference between

host switching and true broad host range. Finally, we discuss how current methods are biased toward the isolation of narrow host range lineages.

The subsequent two chapters focus on the development of novel methodologies to uncover and understand phage-host interactions. In **Chapter 4**, adsorption sequencing (AdsorpSeq) is introduced as a method to identify phage-host interactions in environmental samples. In AdsorpSeq, we focus on the interaction between phages and their receptor molecules on the host cell surface. The method consists of five main steps. First, phages and isolated bacterial cell envelope fractions are added mixed. Then, an incubation allows the phages to bind to cell surface receptors within the isolated cell envelope fraction. In the third step, the membrane envelopes, and phages potentially bound to them, are separated from unbound phages. Subsequently, bound phage DNA is isolated. In the final step, this isolated DNA is sequenced and analysed. With these five steps, we provide a rapid and easily implementable method to identify phages in a mixture that interact with a given host envelope. After benchmarking AdsorpSeq using model phages, we use it to uncover phage-host interactions in a hospital wastewater sample.

Finally, the research in **chapter 5** explores styrene maleic acid lipid particles (SMALPs) as a potential new tool in studies of phage-host interactions. SMALPs are a recently developed tool for isolating bacterial membrane proteins without disturbing their native environments. This is convenient, because this common class of phage cell surface receptor molecules often degrade if not captured within a membrane. In the current exploration of SMALPs as a tool for phage-host interaction studies, we show that diverse model phages interact with SMALPs produced from their hosts. We show that these interactions result in the ejection of phage DNA, as it would if a phage infects a host cell. Furthermore, the binding of phages to SMALPs only occurs when using SMALPs from the host species and when the receptor of the phage is present. We conclude that, although there are some caveats, SMALPs are potentially useful if applied to studies of various aspects of phage-host interaction studies.

Together, this thesis provides insights into the myriad of ways which bacteriophages have evolved to interact with their hosts, including an in-depth study into the tail fibres of a widespread human gut phage. It additionally provides two new methodologies with which phage-host interactions can be uncovered.

Publications

1. **De Jonge, P.A.**; Nobrega, F.L.; Brouns, S.J.J.; Dutilh, B.E.; Molecular and evolutionary determinants of bacteriophage host range; Trends in Microbiology; 2018.

2. Nobrega, F.L.; Vlot, M.; **De Jonge, P.A.**, *et al*; Targeting mechanisms of tailed bacteriophages; Nature Reviews Microbiology; 2018.

3. **De Jonge, P.A.**; Von Meijenfeldt, F.A.B.; Van Rooijen, L.E.; Brouns, S.J.J.; Dutilh, B.E.; Evolution of BACON domain tandem repeats in crAssphage and novel gut bacteriophage lineages; Viruses; 2019.

4. Edwards, R; Vega, A.; …; **De Jonge, P.A.**; *et al*; Global phylogeography and ancient evolution of the widespread human gut virus crAssphage; Nature Microbiology; 2019.

5. **De Jonge, P.A.**; von Meijenfeldt F.A.B.; Costa, A.R.; Nobrega, F.L.; Brouns, S.J.J.; Dutilh, B.E.; Adsorption sequencing as a rapid method to link environmental bacteriophages to hosts; iScience; 2020.

6. **De Jonge, P.A.**; Sibinga D.S.S.; Boright O.A.; Costa, A.R.; Nobrega, F.L.; Brouns, S.J.J.; Dutilh, B.E.; Development of styrene maleic acid lipid particles (SMALPs) as a tool for studies of phage-host interactions; *manuscript under review*.

Curriculum vitæ

Patrick Alexander de Jonge

15-07-1989		Geboren in Arnhem *Born in Arnhem, the Netherlands*
2001	2005	Middelbare school (VMBO-t) *High school* 't Rhedens lyceum, Dieren
2005	2009	Middelbaar beroepsonderwijs (MBO) laboratoriumtechnieken *Vocational education in laboratory techniques* ROC Rijn IJssel, Arnhem
2009	2012	Hoger beroepsonderwijs (HBO) life sciences *Bachelor of Applied Science (BASc) in life sciences* Hogeschool van Arnhem en Nijmegen, Nijmegen
2012	2014	Masteropleiding biologie *Master of science (MSc) in biology* Radboud Universiteit, Nijmegen
2014	2015	Scientific curator ResearchAnt, Rotterdam
2015	2020	PhD kandidaat theoretische biologie en bioinformatica & departement voor bionanowetenschappen *PhD candidate theoretical biology and bioinformatics & department of bionanoscience* Universiteit Utrecht, Utrecht Technische Universiteit Delft, Delft

Acknowledgements

Here I am at last, at the shores of the sea. The date at which I defend my thesis, successfully or otherwise, is five days short of the fifth anniversary at which I started my PhD. In those five years, I have worked at three different Dutch universities (four if you count my current postdoctoral research at the Amsterdam UMC). At each of these universities, I was fortunate enough to meet many great researchers and make many great friends. Although thanking everyone to their full worth would take many pages, here I will try to thank as many people as I can for as much as I can. Knowing my own penchant for forgetfulness, I will likely omit some who deserve to be mentioned here. To those I here belatedly extend my thanks.

First and foremost, I should thank my two supervisors. **Bas,** thank you for the opportunity that you gave me five years ago. When I started my PhD, I had never worked with phages, nor had I ever done any bioinformatics. It is in no small part thanks to your tutoring that I can now continue my research career in phage bioinformatics. Additionally, I feel that without your advice I wouldn't be able to string together more than two words in a scientific text. I hope to become as enthusiastic and thorough of a researcher as you, and I hope that I too will discover an important new phage family so that I can give it a vaguely dirty name. **Stan,** thank you for the hospitality in giving me a place to do my work. Without my time in your research group I would not have developed into the independent researcher that I am today. You allowed me to find my own direction in my research, and to develop my skill set in an organic way. When I joined your lab in Wageningen five years ago, I could not have foreseen where it would take me, both professionally and physically. I am happy to have had a part in building your new research group in Delft. Also, thanks to you my mom got to see me on tv putting samples into a big expensive electron microscope. That was pretty cool too. Next, my promoter. **Berend**, despite the physical distance during much of my PhD, you always showed interest in my work. Thank you for this, and for giving me the support when I needed it most. Then, I would like to extend a special thanks to the members of my committee: **Hilde Herrema, Colin Hill, Martijn Huijnen, Rob Willems,** and **Sasha Zhernakova,** for giving their precious time to reading my thesis (and in one case for giving me a job).

As stated above, I have been a part of three research departments in separate Dutch universities over the course of my PhD. Let me next thank those people spread out across Dutch academia with whom I had the pleasure to work.

I will start with my two paranymphs. They share many things, both now and hopefully long into the future. Among these is the trauma of hearing half a decade of my terrible, terrible jokes. My heartfelt congratulations for surviving them. **Jochem,** our annual Easter hiking holidays to South, West, East and North were among the best times of my PhD. Perhaps someday we shall again find ourselves sitting down beside a deserted path on cliffs overlooking the sea, on top of a snow-capped mountain eating local cheese, or in a train riding though fog whilst playing cards. You are a great intellectual sparring partner, a rock on which I built my experiments, and a true friend through thick and thin. I will never be able to repay you for your help and support, and I am thankful to have you as a friend. Even though you're from Nijmegen (shiver). **Becca,** whenever I needed to rant on things great and small, you were always there. More important, after listening to a rant, you were never afraid to call me out on my bullshit. You are the life of whatever party you're at, and, being the introvert that I am, on more than one occasion I have been glad that you dragged me along with you. Thank you for giving me a pat on the back when I needed it, and I hope that I gained some of your joy, enthusiasm, and kind-heartedness.

To the members of the Brounslab, present and past. **Rita;** it was great to have all the late-afternoon chats in the office, from personal to philosophical to (very) silly. Thank you for always listening to my experimental misadventures. Much of the work in this thesis would not have succeeded without your insight and advice. You're a great researcher, and the best supervisor anyone could wish for. **Franklin,** when I started my project you were my personal phage encyclopaedia, and as my personal job recruiter you found me a position when my projects had ended. Thank you for five years of advice on phage biology and five (and counting) co-authored publications. You're going to be a great PI over in the UK. A gracias to **Christian,** was always the disarmingly optimistic and upbeat note in the office. All the best with those last few steps to your own defence! **Anna,** thank you for keeping the lab actually running and clean. On the topic of keeping the lab running, also a big thank you to **Teunke,** who has been a calm intermitted presence since the first day of my PhD. Further gratitude goes to **Seb** for helping me on the road, whether it was the road to protein purification or to Hannover; to **Tobal** for keeping me from being the only nerd in the group; and to **Boris** for filtering wastewater so I didn't have to. Finally, I thank the students that have aided me in my project **Oliver,** who

SMALPed many a bacterium while quoting the best old school hip hop, and **Dieuwke,** who stayed positive through many hours of fruitless searching TEM grids.

Next to BN, staring with the other two guys in the threesome. **Mehran (from Tehran),** my Persian philosopher roomie during the best years of both of our PhDs. Thank you for all the good discussions that made me examine my opinions, and also for the ones that got quite irreverent quite fast. It is comforting to know that you have now found a true calling in teaching. **Johannes (not from Tehran),** with whom I crossed swords many a time. You are the most honest and down-to-earth person I know, and many could learn from you. I hope that your studies will bring you all that you're looking for. Thank you to calm and collected **Nicole,** always happy **Alicia,** great orator **Roland,** and great charmer **Jonas** for the café card games. The in the wider BN department, my gratitude to **Carsten, David, Lisa, Helena, Mathia, Da, Duco, Benjamin, Fede, Louis, Sam, Anthony, Victorija, Fayezeh,** and all other members, both past and present, with whom I got to share drinks and laughs.

Bastiaan, thanks for all the crash courses in bioinformatics, for helping me even as you yourself were overworked and out of time in your own PhD, and for the chats over microwaved dinners. Further thanks to **Nikos, Laura, Bram, Tina, Eva, Hilje, Juliane, Peter, Julian, Laurens, Richard,** and all the other people at the TBB department. Next, although it is already so long ago, I extent my gratitude to **Tim, Marnix, Yifan, Wen, Prarthana,** and all the people of the laboratory of microbiology in Wageningen with whom I shared the first year of my PhD. Finally, I'd like to thank **Mike** and **Dan** for broadening my horizons and providing endless entertainment during the long hours in the lab.

Last but certainly not least, I'd like to thank those that are closest to my heart. First **Lucy**, who is the very closest. You are the best thing to happen to me during the last five years, and I do not know how I would've gotten through it without your support. I am looking forward to our life together. Aku cinta kamu, sayang. Aan mijn zussen: **Suzanne en Yolande,** mijn emotionele steun in toeverlaat. Iedereen zou zich zulke zusters wensen, die altijd willen luisteren en ondersteunen. Van winter tot zomer, ochtend tot avond, tralala tot dingdingdong, altijd staan hun voor mij klaar. Ik wordt er haast emotioneel van. De een zal nu ongetwijfeld zeggen: 'Ik irriteer me aan al die taalfouten', de ander zal allerlei vragen willen snorren (die ze maar tot later laat scheren). Ook dank aan mijn verhuisspecialist, **Remon,** die altijd bereid is om zijn familie te helpen, en mijn sausspecialist **Michel,** de stille kracht in de familie van wie ik nog

iets kan leren over stressbestendigheid. Aan **Julia, Levi, en Finn,** hoewel jullie nog te klein zijn om dit te lezen, dank dat ze zo af en toe de stress van promoveren doen vergeten. **Geert, Greet, Marinus, en Mien,** dank ik voor alles wat zei voor mij betekend hebben, hoewel ze dit niet meet lezen zullen. **Pap en Mam;** het is nu achttien jaar geleden dat ik als middelbare scholier met suboptimale rapportcijfers thuiskwam. Jullie hebben er toen, en sindsdien, voor gestreden om mij naar mijn volle kunnen te laten presteren. Dit proefschrift is het bewijs van jullie gelijk, dit werk zou hier niet liggen zonder jullie.