Chapter 1: Introduction

1.1 General overview of bacteriophages

1.1.1 The life cycle of bacteriophages

Viruses are the most numerous biological entities on Earth with an estimated population of 10^{31} particles (Brüssow and Hendrix, 2002). Bacteriophages or phages for short, are viruses that infect and replicates within bacteria. Their life cycle beings with the injection of their genome into the cytoplasm of a bacterium followed by either the lytic or lysogenic cycle (Figure 1.1). During the lytic cycle, the cell's metabolism is immediately taken over to replicate the phage DNA and start the synthesis of phage proteins required for the assembly of new viral particles. The cycle finishes when phage lytic enzymes destroy the cell wall and newly formed phages are released from the bacterium (Young, 1992). During lysogeny, the phage genome is either integrated in the bacterium genome or kept as a circular replicon in the bacterial cytoplasm (Lwoff, 1953). At this stage the bacterium is not killed and the carried phage genome is referred to as a prophage while the bacterium becomes a lysogen. Lysogens are able to pass their prophages to daughter cells, however the prophage can be 'awakened' at a future generation and enter the lytic cycle. Phages that exclusively rely on the lytic cycle are called virulent, whereas phages able to enter the lysogenic cycle are called temperate.

Besides the lytic and lysogenic cycles, there are other less studied outcomes of a phage infection. One is displayed by the M13 phage which is able to replicate and generate virions without killing its host (Loh et al., 2019). Another route is when phages are carried inside bacteria but do not integrate or proliferate (pseudolysogeny) (Ripp and Miller, 1997). These phages are inactive in some sense and are asymmetrically segregated upon subsequent divisions. Finally, phages can also accumulate deleterious mutations when integrated in the host genome and as a consequence cannot longer enter the lytic cycle. These defective prophages usually are further degraded, however sometimes a subset of their genes can be beneficial for the host and are conserved (phage domestication) (Bobay et al., 2014).

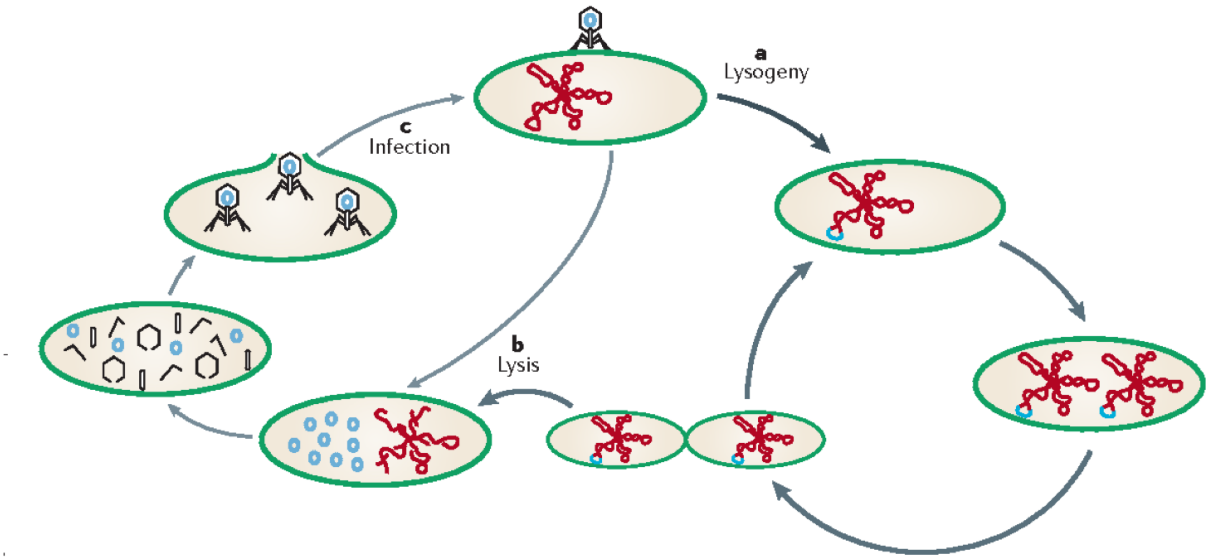


Figure 1.1. The lifecycles of bacteriophages. Lytic and lysogenic are the two main outcomes of a bacterial phage infection. In the former, a phage starts replicating its genome immediately along with the synthesis of phage proteins, ultimately lysing the host and releasing all the newly assembled phages. In the latter, the phage genome integrates into the bacterial genome or is kept as a circular replicon in the bacterial cytoplasm while is passively passed to daughter cells. Sourced from (Reyes et al., 2012)

1.1.2 The outstanding diversity of bacteriophages

Phages can have a DNA or RNA genome (Figure 1.2). However, by far, the most studied phages are those with a linear double stranded DNA (dsDNA) genome. This group is referred to as the *Caudovirales* order and traditionally have been composed of 3 families, namely *Podoviridae, Siphoviridae, and Myoviridae* (Ackermann, 1998). Although with the revised ICTV virus taxonomy from 2019, the *Caudovirales* are now composed of a total of 9 families. A common thing among the *Caudovirales* is the presence of a tail, which is involved in host recognition, cell wall penetration, and genome ejection into the bacteria. *Myoviridae* phages have contractile long straight tails, *Siphoviridae* phages are characterized by non-contractile long flexible tails, and *Podoviridae* phages possess non-contractile short tails. The genomes of *Caudovirales* can vary from 15 kb to 500 kb and are stored in protein complexes called capsids. During virion assembly, 'scaffolding' proteins provide structure for the correct polymerization of the capsid subunits (major capsid proteins) and a connector protein (portal protein) provides a channel for the translocation of the genome into the capsid. Genome packaging is carried out

by a molecular machine composed of the large and small terminases. Replication of DNA generates head-to-tail concatemers of genome units and the small terminase is involved in recognition of phage DNA while the large terminase cuts the DNA concatemer and starts the translocation of DNA fuelled by ATP hydrolysis (Fokine and Rossmann, 2014).

Other less studied phages include the *Tectiviridae* family which possess a linear dsDNA, the *Microviridae* and the *Inoviridae* families which are characterized by having small (<10 kb) and circular single stranded DNA (ssDNA) genomes, the *Leviviridae* family with small (<5 kb) linear ssRNA genomes, and the *Cystoviridae* with dsRNA genomes (Dion et al., 2020).

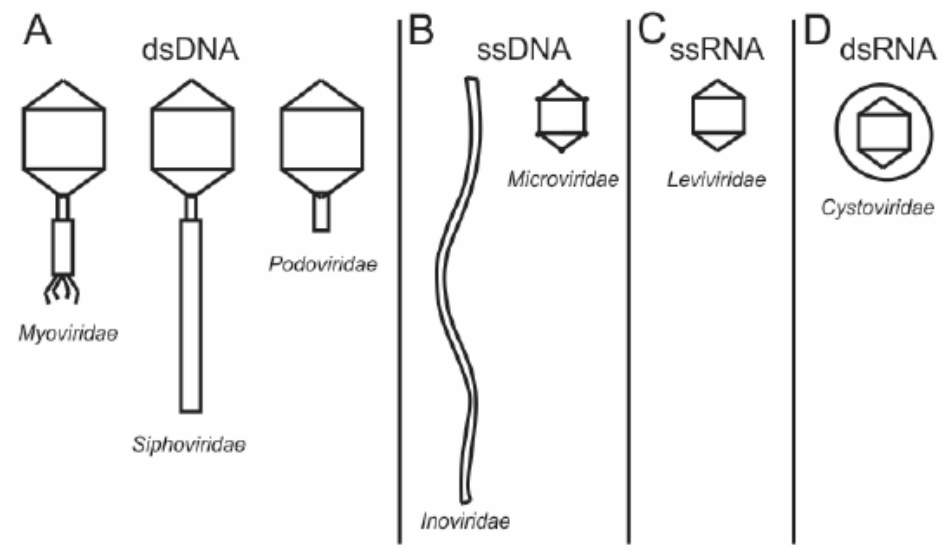


Figure 1.2. The diversity of phages. There is an outstanding phage diversity. Most of the known phages belong to the *Caudovirales* order which traditionally have been divided into 3 families, namely *Myoviridae*, *Siphoviridae*, and *Podoviridae* **(A)**. Other less studied phages include ssDNA phages such as *Inoviridae* and *Microviridae* **(B)**, and phages with an RNA genomes such as *Leviviridae* (ssRNA) **(C)** and *Cystoviridae* (dsRNA) **(D)**. Sourced from (Denton et al., 2013)

1.1.3 Phages: friends or foes of bacteria?

Even though it is tempting to label phages as parasites, they represent a potent force driving ecological functioning and evolutionary change in bacterial communities. A clear example is bacteriophage-mediated horizontal gene transfer, which enhances bacterial adaptive responses to environmental changes (Canchaya et al., 2003). When a prophage undergoes a faulty excision, adjacent chromosomal DNA can end up packaged with the phage genome (specialized transduction) (Morse et al., 1956). A more extreme case can occur when only chromosomal or plasmid DNA is packaged (generalized transduction) (Zinder and Lederberg, 1952).

Phages can also directly increase the fitness of their host. For instance, the viral encoded cI repressor protein which promotes lysogeny of the *E. coli* phage λ, also represses the host gene pckA. This repression in turn causes a decoupling of central metabolism from cellular synthesis, reducing growth rate and may confer a selective advantage in bacteria living in nutrient limited environments (Chen et al., 2005). A more subtle mechanism that phages can use to influence the host phenotype comes from active lysogeny. In this phenomenon, phage excision acts as a regulatory mechanism for expression of bacterial genes without entering the lytic cycle. An example is the phage Φ10403S which its integration disrupts a gene (comK) involved in the escape of its host from the mammalian phagosome. However, when expression of comK is needed, the phage excises and restores the gene function, allowing the survival of its host (Feiner et al., 2015). Other ways bacteria can benefit from phages include the encoding of virulence factors, protection against further phage infection, enhanced biofilm formation, and antibiotic tolerance (Abedon and LeJeune, 2007; Bondy-Denomy et al., 2016; Burmeister et al., 2020; Gödeke et al., 2011).

Co-evolutionary interactions between phages and bacteria also shape the phenotype of bacterial communities. In an effort to prevent successful phage infections, bacteria often mutate and differentially express receptor proteins exploited by phages (Hyman and Abedon, 2010), produce cell surface polysaccharides (Fernandes and São-José, 2018), and can even increase their mutation rate to boost adaptation (Morgan et al., 2010).

1.1.4 The arms-race between phage and bacteria

The Red Queen hypothesis postulates that organisms must constantly evolve and adapt against ever-evolving opposing organisms that share the same environment (Leigh Van Valen, 1973). This scenario is particularly pronounced for bacteria given the constant threat of the lytic cycle and the extremely rapid evolution of phages. Thus, bacteria have developed several strategies to prevent successful phage infections, and at the same time, phages have evolved counter-resistance measures (Figure 1.3).

Bacteria can prevent phage adsorption by altering their receptors (e.g. mutation or chemical modification such as glycosylation) (Harvey et al., 2018) or by masking them with exopolysaccharide capsules (Ohshima et al., 1988). A more indirect approach involves the release of outer membrane vesicles (OMVs) with embedded phage receptors. OMVs thus serve as phage decoys and reduce productive infections (Reyes-Robles et al., 2018). However phages can overcome these hurdles by mutating their receptor-binding proteins (RBPs) to recognize the altered receptors (Meyer et al., 2012), encode multiple RBPs (Schwarzer et al., 2012), or even producing depolymerases to expose a hidden receptor (Fernandes and São-José, 2018).

Even if phages breach extracellular defence mechanisms, bacteria still can counter phages by using intracellular defence systems. Restriction-modification (RM) systems work by cleaving the phage genome upon injection (Oliveira et al., 2014). This is carried out by a restriction endonuclease (R) which recognizes unmethylated phage DNA, while the host DNA remains intact due to methyl modifications by the associated methyltransferase (M). The phage growth limitation (Pgl) system is similar to the RM system except that phages become methylated only after completing the infection cycle (Sumby and Smith, 2002). In a subsequent infection, however, these methylated phages are cleaved upon entry. The DISARM system was recently described and also works by using methylation as an immunity mark, however it provides resistance in the early stages of infection by a yet unknown mechanism (Ofir et al., 2018). Phages have evolved a wide array of strategies to circumvent RM systems (Samson et al., 2013). They can mutate RM sites or modify bases via glycosylation, glucosylation, hydroxymethylation and acetamidation to avoid recognition by the restriction endonuclease. Phages can also activate host methyltransferases or encode their own in order to protect their genome from restriction. Other examples include the Dar system of coliphage P1 which

reduces DNA degradation by interfering with the activity of type I restriction endonucleases and the Ocr protein of coliphage T7 which binds and sequesters the EcoKI endonuclease.

A third type of defence is the CRISPR/cas system which represents a form of adaptive immunity. When a phage infects a bacteria, small fragments of the virus (spacers) are acquired by bacteria. Later on, spacers are transcribed and used as specific probes to recognize phage DNA sequences (protospacers) which leads to degradation by the Cas endonuclease (Barrangou et al., 2007). Phages, on the other hand, can mutate protospacers or modify their bases to avoid recognition by the Cas protein (Paez-Espino et al., 2015), however sometimes escape mutations can lead to phage fitness defects. Anti-CRISPR (Acr) proteins provide a way to overcome this risk by blocking the activity of CRISPR-Cas systems and they do so by mostly interacting with Cas proteins (Bondy-Denomy et al., 2013). As an idea of the complexity of phage/bacteria interactions, CRISPR-cas systems can also be encoded by phages, which can be used to evade host innate immunity (Bondy-Denomy et al., 2013).

In contrast to previous defence systems which focus on protecting individual hosts, abortive infection (Abi) systems act at the population level. They are characterized by allowing phage entry but then the cell host dies in an "altruistic" fashion to severely limit the release of phages and prevent a phage epidemic in the bacterial population (Chopin et al., 2005). Some Abi systems work by exploiting toxin-antitoxin mechanisms. For instance, RnlA is a toxin with endoribonuclease activity which is neutralized by the RnlB antitoxin. Whereas RnlA is a stable protein, RnlB is quickly degraded and thus it needs to be constantly synthesized. However, infection by the T4 phage rapidly shuts off *E. coli* gene expression, resulting in the disappearance of RnlA and allowing RnlB to cause cell death (Naka et al., 2017). Some counter-measures to avoid Abi systems include evolving alternative antitoxins (Otsuka and Yonesaki, 2012), acquiring native antitoxins by recombination with the host (Blower et al., 2012), producing proteins that prevent the degradation of the antitoxin, and directly inhibiting toxins (Alawneh et al., 2016).

Finally, the phage-inducible chromosomal islands (PICIs) are phage parasites that can affect phages by disrupting phage particle assembly and DNA packaging (assembly interference) (Seed, 2015). The best studied members of PICIs are the Staphylococcus aureus pathogenicity islands (SaPIs), which cause mature phage particles to package SaPI DNA rather than phage DNA.

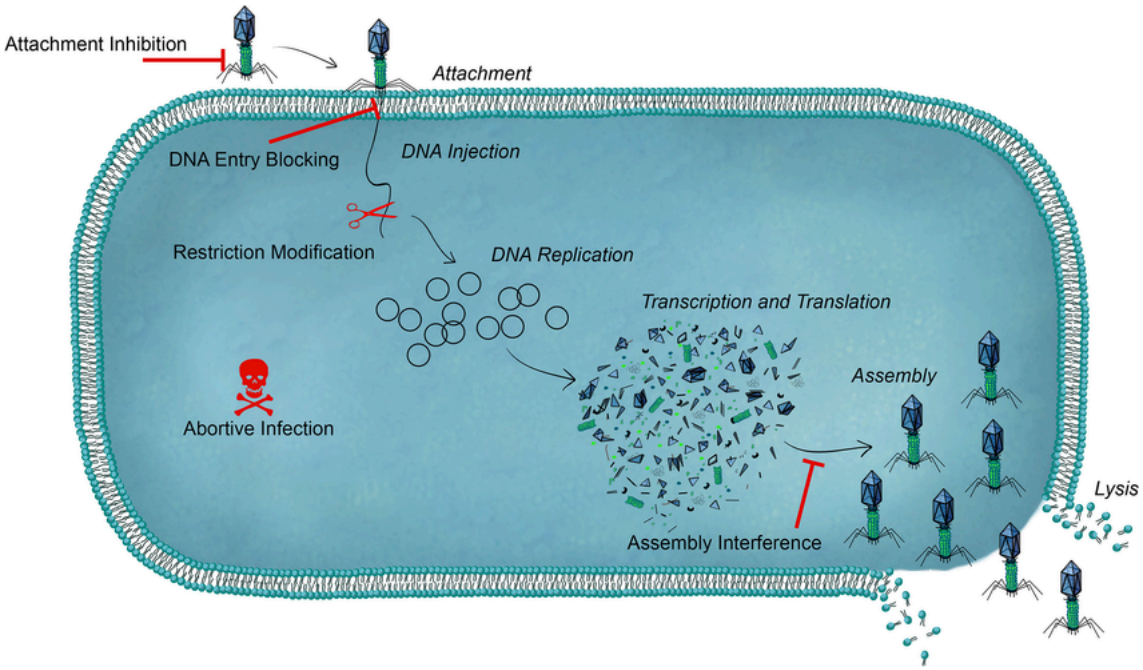


Figure 1.3. Bacterial anti-phage defences. Bacteria have acquired an arsenal of strategies to interfere with phage infections. These defence systems can act by preventing phage attachment and DNA injection, degrading phage DNA by restriction or CRISPR systems, abortive infection, among others. Sourced from (Seed, 2015)

1.1.5 Evolutionary phage-host dynamics

While the previous section highlighted the mechanisms of resistance and counter-resistance, now we review how these strategies vary over time. Two main models have been proposed to explain the dynamics of resistance and counter-resistance. The arms race dynamics model posits that phages select for resistant hosts, which in turn apply selective pressure for phage mutations that restore infectivity, and the cycle repeats. However, coevolutionary experimental studies have shown that the arms race between viruses and bacteria does not continue indefinitely (Hall et al., 2011). One explanation is related to metabolic constraints associated with phage resistance. For instance, if a viral receptor is also a nutrient uptake protein and a resistance conferring mutation impairs nutrient acquisition.

The fluctuating selection model on the other hand, proposes that as the abundance of a fast-growing susceptible host increases, so does the likelihood of encountering a phage, resulting

in increased host mortality and allows for slow-growing resistant bacteria to become majority. However, as the number of phages decreases due to the lack of susceptible hosts, the resistance conferring mutation starts to lose advantage, letting the susceptible fast-growing bacteria to dominate the population and the cycle starts again (Avrani et al., 2012).

1.1.6 Predator-prey dynamics

Whenever we have a predator and a prey interacting, an interesting question arises: how will bacteria and phage populations vary over time? In phage-bacteria interactions two main models have been put forward to explain their dynamics. The first one is the "Kill-the-Winner" model (Thingstad, 2000). This model is based on the assumption that the likelihood of phages killing bacteria is proportional to the relative abundance of the host and mathematically has been approximated with the Lotka-Volterra equations. This way, high levels of bacterial diversity are maintained as overgrown bacteria will be killed by their phages. A second model is "Piggy-back-the-Winner" and it posits that when a host is abundant and growing rapidly, temperate phages will prefer to enter the lysogenic cycle. In addition to replicating "for free" (due to the fast growing rate of its host), they can provide defence against other phages by super infection immunity (Knowles et al., 2016).

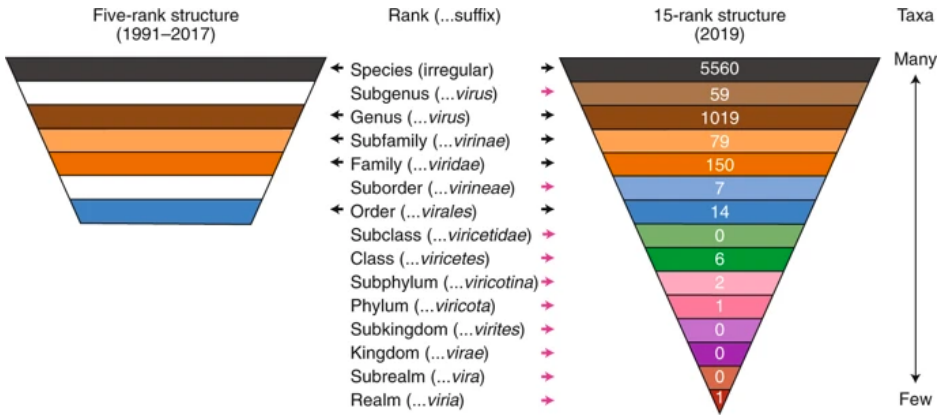
1.1.7 Taxonomy and the recent explosion of phage diversity

The taxonomy of phages is established by the International Committee on the Taxonomy of Viruses (ICTV) which published its first report in 1971. (Adriaenssens and Brister, 2017) Initial classification efforts were based mainly on phage morphology (facilitated by electron microscopy observations) and nucleic acid content, which have been the major criterion for classification at the family taxonomic rank. For many years, most of the phages discovered were categorized to belong to one of the 3 traditional *Caudovirales* families, namely *Podoviridae*, *Siphoviridae*, and *Myoviridae*. However, grouping at lower taxonomic levels such as genus and subfamily was rarely addressed. Demarcation of species in phages is currently set at 95% nucleotide identity, constrained to low levels of genome re-arrangements. In the case of genus, nucleotide identity can drop to 50% as long as the group shares a set of cohesive features such as average genome length, presence of signature genes, average number of tRNAs, etc. Recently, the ICTV has allowed a 15-rank classification which aims to

accommodate the entire spectrum of genetic divergence in the virosphere (Gorbalenya et al., 2020) (Figure 1.4A). This expanded classification matches better the Linnaean taxonomic system. In line with this development, a proposed megataxonomy for all viruses was published this year (Koonin et al., 2020). With this taxonomy current known phages can be placed into other higher orders, for instance, the *Caudovirales* belong to the class *Caudoviricetes*, phylum *Uroviricota*, Kingdom *Heunggongvirae*, and realm *Duplodnaviria*.

With the advent of high-throughput sequencing and metagenomics, there was an explosion on the number of novel phages discovered (Figure 1.4B). With the vast majority of these newly discovered phages only known by sequence, most of them remained unclassified. In an effort to counter this classification issue, several alternative classification schemes were proposed which were based only on sequence information such as the phage proteomic tree, gene-sharing networks, and kmer-based grouping. Proposals to incorporate the vast number of phages discovered by metagenomics into current phage taxonomy are now being considered by the ICTV (Simmonds et al., 2017).

A



B

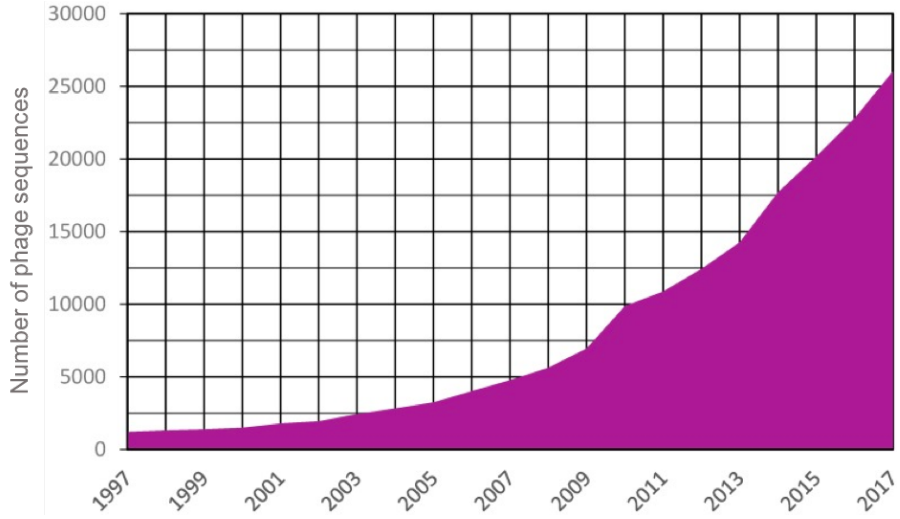


Figure 1.4. Taxonomy of phages. A) The highest taxonomy rank to classify phages was "order". Recently the ICTV incorporated a 15-rank classification which aims to accommodate the entire spectrum of genetic divergence in the virosphere. Sourced from (Gorbalenya et al., 2020). **B)** The number of discovered phage sequences deposited on Genbank across the years was fuelled by high-throughput sequencing and metagenomics. Unfortunately, the majority of sequences remained unclassified. Adapted from (Adriaenssens and Brister, 2017)

1.1.8 Prediction of phages from metagenomic sequences

As mentioned in the previous section, the recent explosion of discovered phage diversity has been fuelled by the mining of metagenomic sequences. A common strategy to identify phages involves the comparison of proteins in the query DNA to a reference database of known phage proteins (Roux et al., 2015). However, this similarity approach is limited to mainly find phages related to the ones in the database, and thus falls short when mining environments with a high level of novel phage diversity. The similarity approach can be improved by the use of Hidden Markov Models, as they are suitable to detect more similarity between novel and known phage proteins.

Another strategy involves the detection of "viral-like" genomic features, such as GC skew, protein length and transcription strand directionality. The use of kmer profiles has also been exploited to differentiate phages from bacterial DNA (Ren et al., 2017).

1.2 Bacteriophages in the human gut

1.2.1 Discovery and isolation of faecal VLPs

Phages in the gut were discovered in 1917 by d'Herelle when he reported "an invisible microbe with antagonistic properties against the Shiga bacillus" in stools from individuals convalescent from bacillary dysentery (D'Herelle, 2007). However, it was not until recently, that more research started to focus on gut phages. In part because of the increased awareness of the gut microbiota in human health, and because gut phages often prey on bacterial hosts which traditionally have been very challenging to cultivate (strict anaerobes) (Browne et al., 2016). Even though now it's technically possible to culture a large number of anaerobic bacteria from the gut, a wealth of information about gut phages has come from the analysis of viral nucleic acids extracted from human faeces. A common procedure, involves the use of 0.2 or 0.45 um filtered faecal samples to greatly reduce non-viral contamination, followed by several physical and enzymatic steps that remove prokaryotic and eukaryotic material (Shkoporov et al., 2018a). The resultant supernatant is enriched in virions, or viral like particles (VLPs) which are then digested to release and sequence the viral nucleic acids. A disadvantage is that VLPs represent only phages that are undergoing the lytic cycle, and thus inactive prophages at the moment of VLP extraction are missed.

1.2.2 Taxonomy of gut phages

Microscopic studies of VLPs and their nucleic acids has shown that the gut phageome is dominated by members of the *Caudovirales* (Hoyles et al., 2014) (Figure 1.5). Other studies have also detected other families such as *Microviridae* and *Inoviridae* (Kim et al., 2011). RNA phages, although present in faeces, are thought to be rare. In addition, giant phages with a genome size > 540 kb in length have been detected in human faeces from Bangladesh. These phages which were assigned a *Prevotella* host, are thought to be enriched in the gut microbiome of individuals who consume non-Western diets (Devoto et al., 2019).

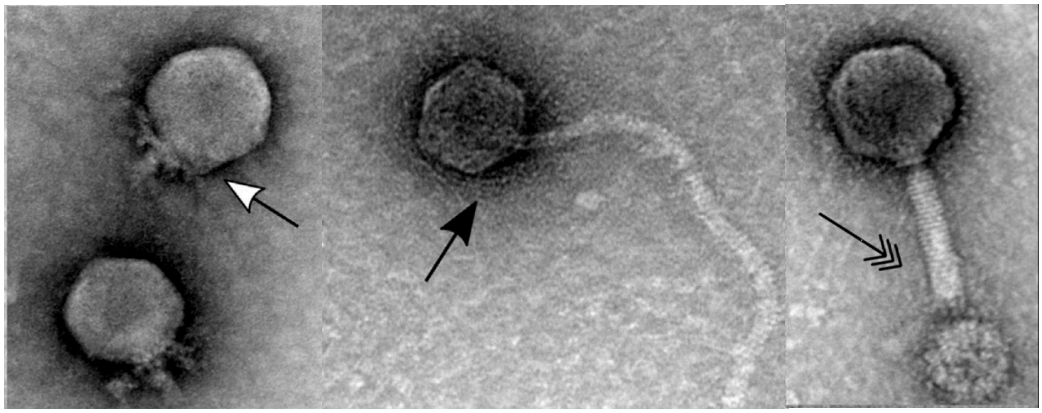


Figure 1.5. Main bacteriophage morphological types detected in a faecal sample. The main phages identified in human faeces belong to the *Caudovirales* order. Here, highlighted from left to right the *Podoviridae*, *Siphoviridae*, and *Myoviridae*. Adapted from (Shkoporov and Hill, 2019)

1.2.3 The case of the crAssphage

The most famous human gut phage is the crAssphage, which was first reported in 2014 and its genome was assembled purely from metagenomic reads (thus the name CRoss-ASSembly) (Dutilh et al., 2014). This phage which is highly prevalent in Western cohorts and can represent up to 90% of the total reads from a single virome, went undetected for years because it represented a completely novel clade of phages. It was later discovered that crAssphage was a member of an expansive bacteriophage family named "crAss-like" which consisted of 4 subfamilies and 10 genera (Guerin et al., 2018). The original member crAssphage belongs to genus I, and it's often referred to as p-crAssphage (prototypical). Its match with CRISPR spacers, the presence of a *Bacteroides* protein domain (BACON) in its genome, and bacterial abundance correlation experiments suggest that p-crAssphage infects a *Bacteroides* species, however its exact host remains elusive to date. On the other hand, a member of genus VI was isolated in the laboratory from *Bacteroides intestinalis* (Shkoporov et al., 2018b)

1.2.4 Phage dynamics in the human gut

It's thought that lysogeny is the predominant lifestyle of phages in the human gut. This is based on the high number of commensal bacteria harbouring prophages (Kim and Bae, 2018), the abundant genes associated with lysogeny in metagenomic studies, the long-term stability of the gut phageome, and low mutation rate over time in temperate-like contigs. (Minot et al., 2013; Reyes et al., 2010a). In addition, some studies have reported relatively low counts of viral particles with 10^9-10^{10} particles per gram of faeces compared to 10^{11}-10^{12} bacteria. Even adjusting for inefficiencies in the purification process, the number of particles still would be in a range of 10^{10}-10^{12} particles per gram of faeces. When taking into account these estimates, the virus to microbe ratio (VMR) in the gut is significantly lower compared to other microbial communities (Manrique et al., 2017).

In addition to the low VMR observed in the gut, the absence of abundance oscillatory patterns of phages and gut bacteria (which are indicative of a kill-the-winner scenario) (Minot et al., 2011), along with the high rate of suggestive lysogeny in the gut, has led to the proposal that Piggyback-the-Winner (PtW) dynamics predominate in the human gut.

However, dynamics between phage and bacteria may deviate from PtW depending on the distance from the intestinal mucus (Figure 1.6). It has been observed that the VMR is in average four times higher in metazoan-associated mucosal surfaces when compared with the surrounding environment (Silveira and Rohwer, 2016). Given that the VMR is positively correlated with the proximity to the intestinal mucus, it has been proposed that lysogeny is favoured at the top mucosal layer, while a lytic lifestyle predominates in the bacteria-sparse intermediary layers (Silveira and Rohwer, 2016). The bacteriophage adherence to mucus (BAM) postulates that metazoan mucosal surfaces and phage co-evolve to maintain phage adherence which limits microbial colonization of the inner layers.

In the case of the infant microbiome, PtW dynamics may not predominate, as there is instability caused by a marked contraction of phage diversity during the first 2 years of life. This type of dynamics aligns better with a kill-the-winner scenario as predicted by the Lotka-Volterra model, which predicts a decay of predators when there is scarce prey (Lim et al., 2015).

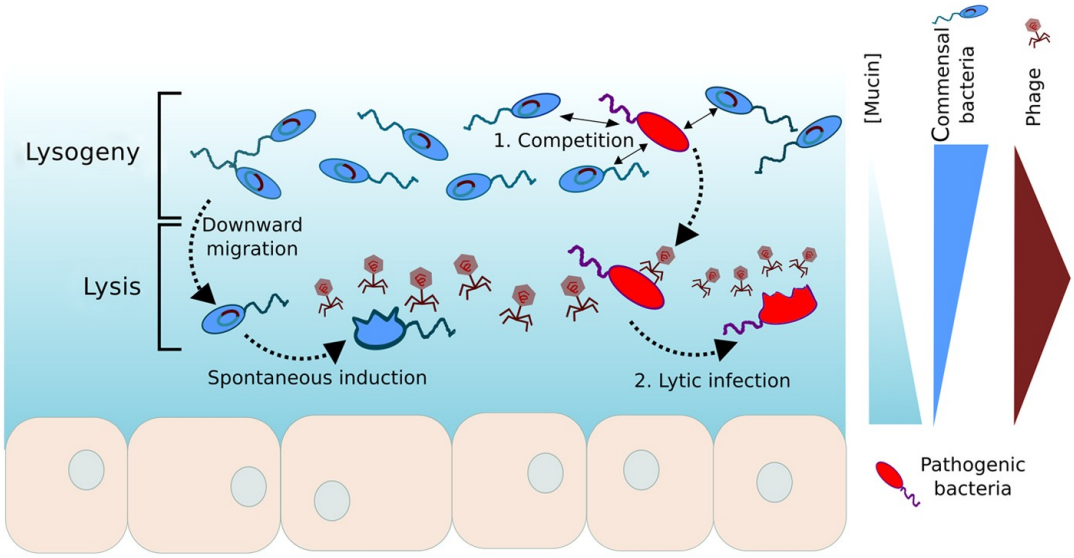


Figure 1.6. Phage dynamics in the human gut. Lysogeny is proposed to be the most prevalent phage cycle in the densely populated human gut (piggyback-the-winner). However, as bacteria move across the intestinal mucus, the lytic cycle is favoured over lysogeny. The bacteriophage adherence to mucus (BAM) postulates that metazoan mucosal surfaces and phage co-evolve to maintain phage adherence which limits microbial colonization of the inner layers. Sourced from (Silveira and Rohwer, 2016)

1.2.5 From lysogeny to the lytic cycle in the gut

Given the high-level of suspected lysogeny in the gut, a key question is whether prophages are active or have become remnants of past phage infections. While there is no a comprehensive study that has evaluated the active fraction of prophages in the gut, a significant proportion of prophages detected by genomic analyses are active (Cornuault et al., 2018; Krupovic and Forterre, 2011; Lugli et al., 2016). In general, phages enter the lytic cycle when they sense a stressor (e.g. activation of the SOS response), it's a survival mechanism that allows them to "abandon a sinking ship". In that regard, induction of gut prophages has been observed by antibiotics (Zhang et al., 2000), diet (such as fructose and short chain fatty acids) (Chatterjee and Duerkop, 2019), bile (Kim et al., 2014), and intestinal inflammation (Diard et al., 2017).

1.2.6 Hosts and host ranges of gut phages

Due to the difficulty of culturing anaerobic gut bacteria, the identity of the hosts targeted by gut phages is a crucial but largely unanswered question. Bioinformatically, CRISPR spacers have been used to link gut phages with predicted hosts. For instance, *Adi et al*. assigned 31 phage contigs to 11 bacterial hosts, with 14 of these phages targeting *Bacteroides* and *Parabacteroide*s (Stern et al., 2012). In another study, one third of 180 phage clusters were linked to abundant taxa such as *Faecalibacteriu*m and *Bacteroides* (Shkoporov et al., 2019). Often phages are restricted to infect single bacterial species, however, intestinal phages may be more promiscuous than expected. For instance, Shkoporov et al. found several phages with broad host range (Shkoporov et al., 2019) and a phage infecting *Faecalibacterium prausnitzii* was shown to also infect *Blautia hansenii* which belongs to a different bacterial taxonomic order (Cornuault et al., 2018). In addition, host range expansion has been observed in a mouse model (De Sordi et al., 2017). However, a study that used a viral tag approach which analysed 363 unique host-phage pairings, found no phages that targeted more than one bacterial species (Džunková et al., 2019). Viral tagging involves the labelling of anonymous virions with a fluorochrome and then they are allowed to attach to host cells. Finally, host-phage pairs are separated by FACS and sequenced to identify the host and the virion. On the other hand, a more comprehensive survey of the host range of gut phages by meta3C proximity ligation (6,651 unique host-phage pairs), found that ~31% of gut phages were not restricted to a single species (Marbouty et al., 2020).

1.2.7 Commonly encoded genes by gut phages

Early insights about the biology of gut phage communities came from the analysis of genetic variation in phage contigs derived from human gut metagenomes (Minot et al., 2012). Hotspots of hypervariation were found in genes homologous to the tail-fibre gene of the Bordetella phage BPP-1, which is hypermutagenized by a unique reverse-transcriptase (RT)-based mechanism (Liu et al., 2002). Moreover, most of the hypervariable loci were linked to genes encoding RTs, highlighting the importance of RTs in the generation of genetic variation for some gut phages.

Other genes that have been found in gut phages are proteins bearing domains from the immunoglobulin (Ig) superfamily. Phages with Ig-like domains have been detected in many

environments, particularly those adjacent to mucosal surfaces. Interestingly, *in-vitro* studies have shown that enrichment of phage in mucus occurs via interactions between Ig-like protein domains and mucin glycoproteins (Barr et al., 2013).

1.2.8 Stability, inter- and intra-diversity of the human gut phageome

The human gut phageome can be defined as the aggregate of phages that inhabit an individual's intestine. It has been found that the human gut phageome is highly diverse between individuals, while intrapersonal variation is minimal and stable (Figure 1.7A,B). In a seminal work (Reyes et al., 2010b), Reyes et al. characterized the faecal viromes of four pairs of adult female monozygotic twins and their mothers by sequencing DNA from VLPs. Analysis of beta diversity revealed that despite remarkable inter-personal variations in their viromes, intra-personal diversity was very low, with >95% of virotypes retained within at least one-year period. Importantly, relative abundances showed minimal variation as well. More evidence about the stability of the gut phageome came from a longitudinal study that monthly tracked the gut phageome of 10 individuals over a period of 1 year by VLP shotgun sequencing. This study revealed that despite certain fluctuations over time, the phageome composition was stable at family and contig level (Shkoporov et al., 2019). This stability was mirrored by the bacterial gut composition which remained stable and individual specific. Another study investigated the relationship between the bacterial microbiome and the virome diversity in 21 adult monozygotic twin pairs (Moreno-Gallego et al., 2019). They found that viromes were unique to individuals, as only 2.83% of the total dereplicated viral contigs were detected in at least 50% of the individuals, and 0.1% were present in all individuals. Notably, this study also showed that phages are the dominant viruses in human gut microbiome, as only 6.42% of the contigs were annotated as Eukaryotic viruses.

The composition of the gut phageome can be altered with diet, however at a lesser degree than interpersonal variation (Minot et al., 2011). Importantly, the variation detected was significantly correlated between bacterial and VLP communities, indicating that diet may affect the gut phageome by perturbing the bacterial gut microbiome.

In contrast to adults, the gut phageome from infants has been found to be less stable. The gut of an infant at birth is considered sterile, but its rapid colonization by microbes derived from

the mother and the surrounding environment leads to the colonization by a phage community. From birth to 2 years of age, there is a contraction and shift in the bacteriophage gut composition, which is in stark contrast with the stable microbiome observed in adults. Moreover, richness and diversity of the gut phageome were found to decrease with age (Lim et al., 2015). Another interesting feature of the infant gut phageome is that the *Caudovirales* and *Microviridae* show an inverse correlation in abundance and diversity up to 2 years of life.

Finally, a controversial concept that has emerged in the field is the existence of a core phageome (Figure 1.7C). Despite the high interpersonal variation found in the gut phageome in previous studies, Manrique et al. proposed that there exists a set of shared phages across individuals referred to as the core phageome (Manrique et al., 2016). In this work, 23 bacteriophages were shared in more than one half of 64 healthy individuals around the world. Moreover, this core set of phages was significantly decreased in individuals with gastrointestinal disease such as IBD. However, a more recent report found that no viral population was detected in more than half of 132 healthy individuals. Specifically, only 1% of phages was shared by over 20% of individuals (Gregory et al., 2019).

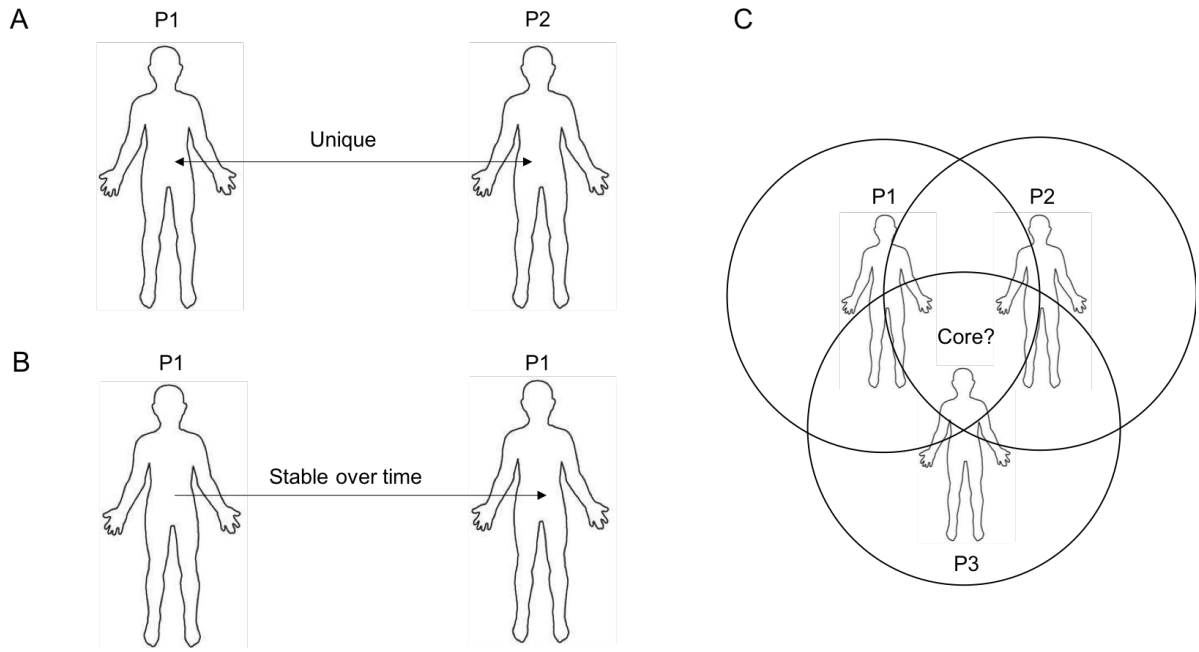


Figure 1.7. Inter- and intra-diversity and stability of the human gut phageome. A) Analysis of phage contigs derived from sequencing faecal viral-like particles (VLPs) has shown that inter-personal variation of the gut phageome is very high among individuals. **B)**

Conversely, the individual gut phageome is stable. **C)** It has been proposed that there is a set of phages shared by a large fraction of individuals, the core phageome. However, this idea is controversial as some studies cannot identify a core phageome.

1.2.9 Gut phages and human disease

Gut phages have been associated with several diseases such as IBD. For instance, it was found that in Crohn's disease and ulcerative colitis the enteric phageome richness increased and that bacterial diversity didn't explain the associated phageome pattern (Norman et al., 2015). However, a subsequent study didn't find evidence of increased phage richness in IBD patients. Instead, it found that healthy controls harboured a stable core of virulent phages that were replaced by temperate phages in Crohn's disease (Clooney et al., 2019).

Another study correlated the increase of strictly lytic virulent lactococcal gut phages with a decrease in Lactococci in Parkinson's disease (PD) patients (Tetz et al., 2018). Lactic acid bacteria are known to produce dopamine and regulate intestinal permeability which are factors implicated in PD pathogenesis. Thus, phages could indirectly contribute to disease by killing beneficial gut bacteria.

Phages can also cause disease by transforming bacteria into pathogens. Certainly, many well-known human diseases are caused by prophage encoded virulence factors such as cholera, diptheria, botulism, and those carrying the Shiga toxin.

1.2.10 Phage therapy

Phages can also be harnessed to treat disease. Shortly after the discovery of phages in 1915, it was realised that they could be used to kill pathogenic bacteria. This idea materialized in 1919 when d'Hérelle first successfully treated several children who were suffering from severe dysentery (Abedon et al., 2011). However, after the discovery of antibiotics, they were disregarded as therapeutic agents particularly in the West (Wittebole et al., 2014). With the rise of antibiotic resistance, there has been a global renewed interest in using phages to treat infections. Unlike antibiotics, phages can be easily mutated to recognize resistant strains, making them very robust to antibiotic resistance; A cocktail of phages can also be used to

mitigate the risk of resistance. In addition, since phages can be very specific to its target strains, there is minimal collateral damage to other bacteria (e.g. gut commensals). Nonetheless, phage therapy also faces some hurdles. For instance, phages can elicit innate and acquired immune responses against them, causing a decrease of their antibiotic activity. The use of temperate phages is inadvisable, given their inherent capacity to the risk of horizontal gene transfer. Phages also contain a large fraction of hypothetical proteins, which could encode proteins that alter bacterial physiology in unexpected ways (Altamirano and Barr, 2019).

Thus, before phages can be deployed as antibiotic agents in different ecosystems such as the human gut, it's necessary to obtain a comprehensive view such as their genomes. Compilation of gut phage genomes could help reveal the function of their genes (e.g. which phages encode virulence factors), identification of the most amenable phages for genetic engineering, their host range, and even assessment of their immunogenicity.

1.3 Thesis aims

The goal of this thesis was to generate critical knowledge about the human gut phageome by harnessing publicly available human gut metagenomes and cultured gut isolates.

Specifically, this thesis aims to:

1) generate the most comprehensive and high-quality database of human gut phage genomes (Chapter 3) ;
2) learn about the functions encoded by gut phages, relevant phage clades, and their bacterial hosts (Chapter 4);
3) investigate global epidemiology patterns of the human gut phageome (Chapter 5).

The objectives relevant to each aim are stated under the introduction of each chapter.