Chapter 4: Function, phylogeny and host assignment of gut phages

4.1 Introduction and aims

Analyses of predicted phage sequences from gut metagenomes have yielded fascinating insights into phage biology, such as the presence of sticky domains - which may facilitate adherence of some phage to the intestinal mucus (Barr et al., 2013) - reverse transcriptases to promote hypervariation (Minot et al., 2012), and proteins with ankyrin domains that may aid bacterial hosts in immune evasion (Jahn et al., 2019). However, previous functions have been inferred from bulk viral fragments, severely limiting the resolution to characterize individual phage genomes.

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. Often phages are restricted to infect single bacterial species, however distantly related gut bacteria have been found to harbour CRISPR spacers that target similar phages (Shkoporov et al., 2019) and almost identical prophages (Cornuault et al., 2018). These results suggest that gut phages may be more promiscuous than expected.

In this chapter, I describe common functions and auxiliary metabolic genes encoded by human gut bacteriophages. I also highlight instances of hypervariable domains which may indicate the presence of phage receptor binding proteins. I then shift the focus to the analysis of two clades of gut phages, namely the Gubaphage and the *Picovirinae* subfamily. The Gubaphage is the viral cluster (VC) with the highest number of GPD predictions after the p-crAssphage, while the *Picovirinae* subfamily was the most common predicted phage taxonomy in GPD. As I will show in Chapter 5, both clades are also highly prevalent across all continents. Finally, host assignment allows me to study patterns of phage diversity across bacterial clades of the human gut and investigate their host range patterns.

The aims of the research presented in this chapter are:

- uncover functions encoded by human gut bacteriophages;
- identify and characterize important phage clades of the human gut;
- carry out host assignment and investigate patterns of phage diversity across gut bacteria.

4.2 Results and discussion

4.2.1 Functions encoded by gut phages

Having a collection of over 142,000 viral genomes from the human gut allowed me to explore the functional patterns of gut bacteriophages at an unprecedented scale. In order to avoid biases due to a large number of highly genetically related viral predictions, I carried out the analysis at the level of VCs and ranked the results by fraction of VCs encoding the predicted functions. In addition, given that prophages are found in GPD predictions, I only considered regions classified as "viral" by checkV (Nayfach et al., 2020) to safeguard against bacterial DNA. I investigated the most ubiquitous KEGG pathways and modules encoded by gut phages (Figure 4.1A). The most frequent KEGG pathways detected were those associated with DNA replication (ko03030), mismatch repair (ko03430), purine and pyrimidine metabolism (ko00230, ko00240), homologous recombination (ko03440), and cysteine and methionine metabolism (ko00270). Although DNA replication, mismatch repair and homologous recombination can be thought of inherent pathways of phages, the last two are an example of auxiliary metabolic genes (AMGs). AMGs augment host metabolism during infection and have a bacterial origin (Breitbart et al., 2007). Inspection of purine and pyrimidine metabolism genes revealed that dUTPases and thymidylate synthases were prominent members of this category. Cellular dUTPases break down dUTP into dUMP and pyrophosphate, while thymidylate synthases convert dUMP into dTTP (Hizi and Herzig, 2015). Since most DNA polymerases can use dUTP instead of dTTP for DNA synthesis, gut phages can minimize the risk of misincorporation of uracil in their genome by lowering the intracellular dUTP/dTTP ratio with dUTPases and thymidylate synthases.

I also found other frequent functions related to the metabolism of sulphur-containing compounds such as assimilatory and dissimilatory sulphate reduction (M00176 and M00596). I decided to specifically search for hits that included the phosphoadenosine phosphosulfate reductase and sulfate adenylyltransferase as both enzymes participate in the reduction of sulfate (Muyzer and Stams, 2008). Sulfate reduction can be harnessed for assimilatory (anabolic) reactions which are involved in the biosynthesis of S-containing amino acids, as well as for dissimilatory pathways (energy generation) which use sulphur instead of molecular oxygen as an electron acceptor. This analysis unveiled 215 VCs that primarily infect *Bacteroides*,

Bacteroides B, *Parabacteroides*, *Prevotella*, *Bacteroides A*, and *Blautia_A*. Phages encoding sulphur metabolism enzymes may seem enigmatic, however dissimilatory reactions could be exploited by phages to ensure sustained energy generation in the gut anaerobic environment. For instance, cyanophages can encode photosynthetic genes in order to boost energy production during the infection stage (Clokie and Mann, 2006). Sulphur metabolism genes have also been found in dsDNA phages from the deep ocean, where it has been hypothesized that they may be involved in supplementing or sustaining sulphur oxidation metabolism in bacteria to ensure continued viral infection and replication (Anantharaman et al., 2014). While the top predicted hosts are not considered sulphur-reducing gut bacteria, it has been shown that *Parabacteroides* and *Bacteroides* isolated from chicken cecum express proteins related to sulfate assimilation. In addition, when dietary carbohydrates are scarce, *Bacteroides thetaiotaomicron* can degrade host glycans (heparin and heparin sulfate) which have variable sulfation patterns. *Prevotella* strain RS2 and *Bacteroides fragilis* are also considered mucin-degrading bacteria (Tailford et al., 2015). Thus, it remains a possibility that as these bacteria can metabolize sulphated compounds, phages could exploit sulphur pathways for their own advantage.

When I was inspecting annotations of individual genomes of GPD phages, I discovered multiple genes annotated as transporters. Therefore, I decided to quantify the most common phage transporters found in GPD (Figure 4.1B). Top hits corresponded to transporters for pantothenate, Zinc, Cobalt, Taurine, Nicotinamide mononucleotide, Nicotinamide riboside, spermidine/putrescine, and potassium.

Nutrient transporters have been identified in other phages. For instance, viral genomes from the North Atlantic Subtropical Gyre can code for the pstS gene which transports phosphate into the host (Warwick-Dugdale et al., 2019). Phosphate is a primary limiting nutrient in marine environments, so phages can benefit their host by coding for phosphate transporters. Certainly, phages isolated from phosphate limited environments have been found to carry more AMGs related to phosphate uptake than those from phosphate replete environments (Kelly et al., 2013). It's known that the human gut is not a homogenous environment but one with nutrients that vary in space and time (gut biogeography) (Donaldson et al., 2016). Thus, the type of transporters coded by phages may depend on nutrients that maximize the chances of survival of their bacterial host at a specific gut niche. In line with this thought, substrates that aid anaerobic respiration may be more common in the most hypoxic areas of the gut such as the

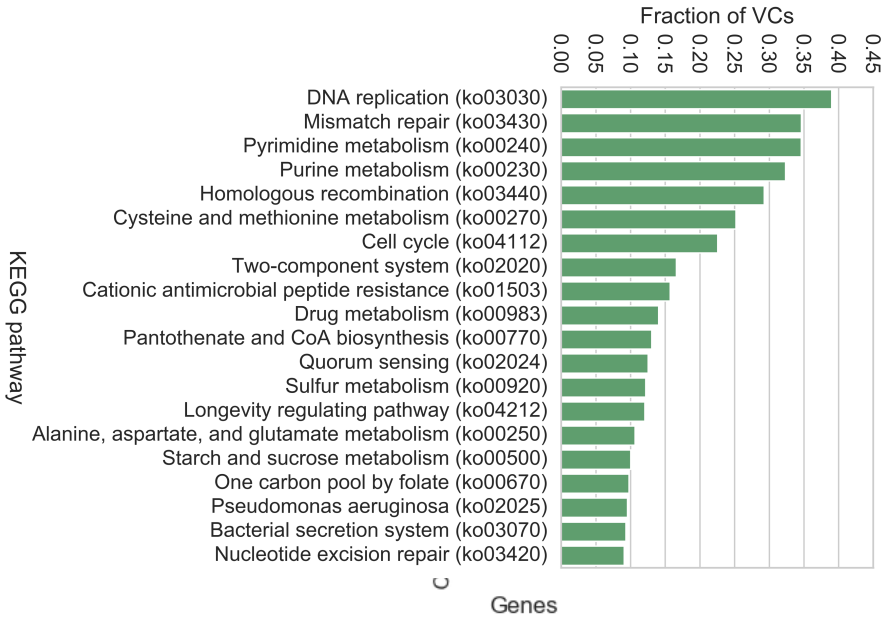
large intestine. For instance, Taurine (a major constituent of bile) can be metabolized into sulfite, enabling anaerobic respiration. Small amounts of bile salts that were not absorbed in the small intestine, may be better harnessed by phages coding for taurine transporters in the hypoxic environment of the large intestine.

I then shifted my attention to investigate the incidence of specific genes previously found in viral metagenomes from human faeces such as reverse transcriptases (Minot et al., 2012) and sticky domains (Barr et al., 2013).

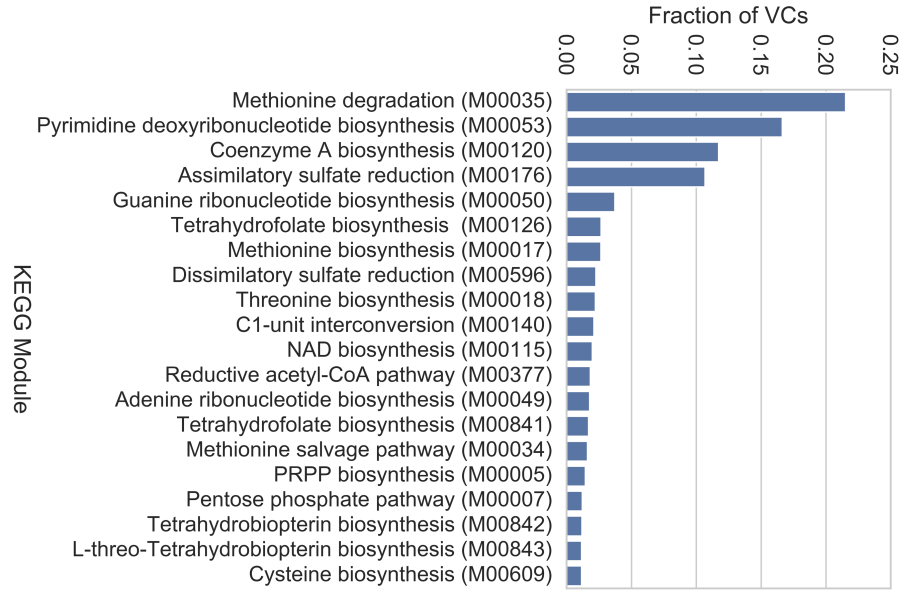
Over 2500 VCs (~12% of all VCs) encode reverse transcriptases (RTs) (Figure 4.1C). RTs in phages have been found to play a role in the generation of sequence diversity in target phage genes such as receptor binding proteins, and thus RTs with that function are called diversity-generating retroelements (Liu et al., 2002). The high incidence of RTs found here contrasts with previous reports that found very low prevalence of DGRs in phages (3 phages in ~600 dsDNA phages from NCBI) (Schillinger and Zingler, 2012). Similarly, When I analysed the incidence of RTs in RefSeq phages, only 0.38% contained them. Recently, it was reported that retrons, which are composed of a RT and a non-coding RNA, can work as an anti-phage defence system (Millman et al., 2020) . It's possible that many RTs carried by gut phages may be involved in defending against other phages, thus providing their host a selective advantage.

I also detected phage genes with adhesive domains (Figure 4.1C). For instance, Immunoglobulin-like (Ig-like) domains which occur frequently on the surface of the *Caudovirales* (Fraser et al., 2006), were found in ~5% of VCs. The Bacteroides-Associated Carbohydrate-Binding Often N-terminal domain (BACON), which has been hypothesized to help phages bind intestinal mucin (de Jonge et al., 2019), was found in 0.88% of VCs. Finally, the collagen triple helix repeat (CTHR) was found in ~8% of VCs. Collagens domains have been suggested to aid in the attachment of phages to *E.* coli (Yu et al., 2014). Sticky domains in phages are often found close to tail genes, and it has been suggested that they may facilitate phage adsorption to its host (Fokine and Rossmann, 2014). In many cases, successful phage infections in the gut are mediated by the correct combination of sticky domains and capsular polysaccharides on the surface of bacteria (Porter et al., 2020).

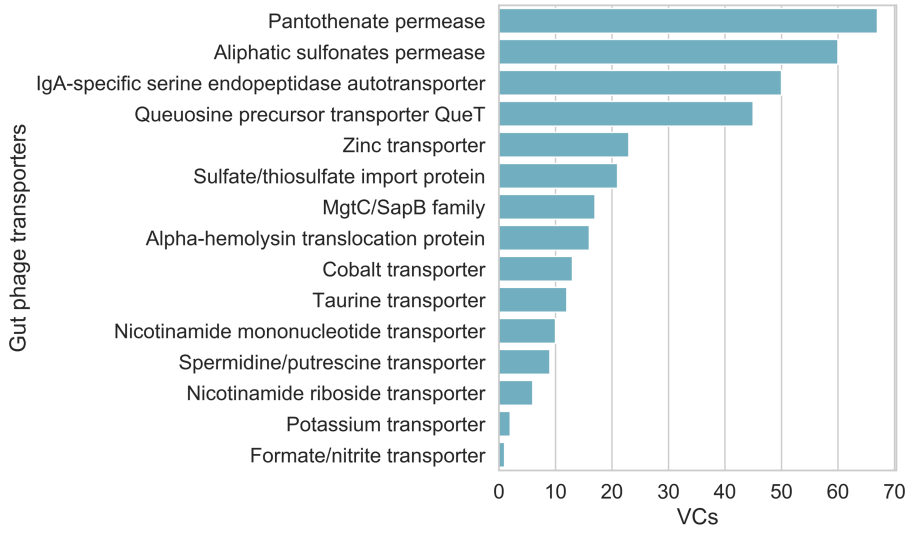
A


Fraction of VCs
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
KEGG pathway
DNA replication (ko03030)
Mismatch repair (ko03430)
Pyrimidine metabolism (ko00240)
Purine metabolism (ko00230)
Homologous recombination (ko03440)
Cysteine and methionine metabolism (ko00270)
Cell cycle (ko04112)
Two-component system (ko02020)
Cationic antimicrobial peptide resistance (ko01503)
Drug metabolism (ko00983)
Pantothenate and CoA biosynthesis (ko00770)
Quorum sensing (ko02024)
Sulfur metabolism (ko00920)
Longevity regulating pathway (ko04212)
Alanine, aspartate, and glutamate metabolism (ko00250)
Starch and sucrose metabolism (ko00500)
One carbon pool by folate (ko00670)
Pseudomonas aeruginosa (ko02025)
Bacterial secretion system (ko03070)
Nucleotide excision repair (ko03420)
Genes

B


Fraction of VCs
0.00
0.05
0.10
0.15
0.20
0.25
KEGG Module
Methionine degradation (M00035)
Pyrimidine deoxyribonucleotide biosynthesis (M00053)
Coenzyme A biosynthesis (M00120)
Assimilatory sulfate reduction (M00176)
Guanine ribonucleotide biosynthesis (M00050)
Tetrahydrofolate biosynthesis (M00126)
Methionine biosynthesis (M00017)
Dissimilatory sulfate reduction (M00596)
Threonine biosynthesis (M00018)
C1-unit interconversion (M00140)
NAD biosynthesis (M00115)
Reductive acetyl-CoA pathway (M00377)
Adenine ribonucleotide biosynthesis (M00049)
Tetrahydrofolate biosynthesis (M00841)
Methionine salvage pathway (M00034)
PRPP biosynthesis (M00005)
Pentose phosphate pathway (M00007)
Tetrahydrobiopterin biosynthesis (M00842)
L-threo-Tetrahydrobiopterin biosynthesis (M00843)
Cysteine biosynthesis (M00609)

C


Gut phage transporters
Pantothenate permease
Aliphatic sulfonates permease
IgA-specific serine endopeptidase autotransporter
Queuosine precursor transporter QueT
Zinc transporter
Sulfate/thiosulfate import protein
MgtC/SapB family
Alpha-hemolysin translocation protein
Cobalt transporter
Taurine transporter
Nicotinamide mononucleotide transporter
Spermidine/putrescine transporter
Nicotinamide riboside transporter
Potassium transporter
Formate/nitrite transporter
0
10
20
30
40
50
60
70
VCs

D

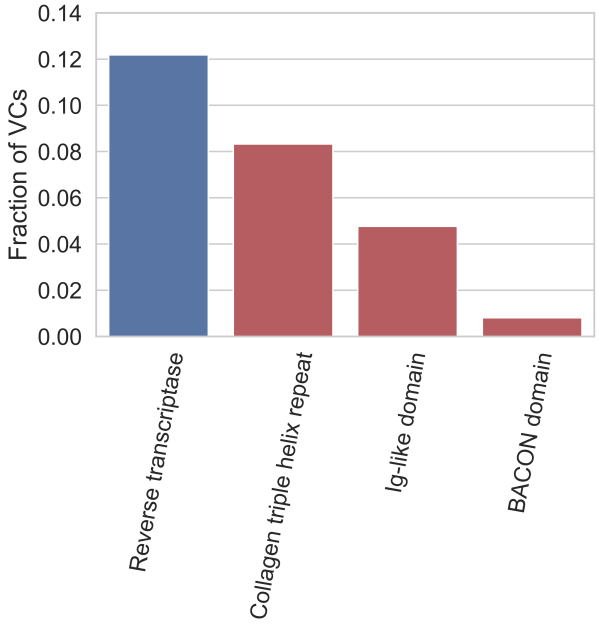


Figure 4.1. Functions encoded by gut phages. **A)** Top functions encoded by gut phages. Common functions included KEGG pathways and modules related to DNA replication and DNA repair. However, I also detected instances of auxiliary metabolic functions such as those involved in nucleotide and sulphur metabolism. **B)** Transporters found in gut phages which may provide a selective advantage to their hosts depending on its intestinal niche. **C)** Reverse transcriptases (RTs) can help phages to generate sequence diversity and potentially act as defence systems against other phages. Sticky domains (red) may facilitate adsorption to hosts and binding to intestinal mucus.

4.2.2 Protein clusters encoded by gut phages

While the functions described above corresponded to curated pathways and targeted searches, I then took a more agnostic approach by analysing the whole proteome of GPD. I clustered all the GPD proteins with the phage RefSeq proteome to understand the functions encoded by the resultant protein clusters (PCs) (Figure 4.2A). After removing singletons I ended up with 172,449 PCs. Top hits included PCs containing proteins involved in the integration of DNA into the host and the maintenance of a lysogenic state (anti-repressor and integrases), DNA processing (single-stranded DNA-binding protein), pore formation for DNA injection (tape-measure protein), DNA packaging into procapsids (terminases), and DNA methylases (defence against host endonucleases). Interestingly, the 11th most common PC (PC_11) which was

encoded by ~8.5% of all VCs could not be clustered with any viral protein from RefSeq. I inferred that this PC encompassed a family of relatively large (median: 259 aa, IQR: 33 aa) single-pass membrane proteins, as they carry a transmembrane region near the N-terminus. Submission of members of PC11 to HHpred (Söding et al., 2005), one of the most sensitive tools for protein homology detection, could not retrieve confident hits. Prediction of the host range of phages carrying proteins that belonged to this PC11, showed that it was mainly found in the Firmicutes phyla. This unknown PC highlights our lack of understanding of 'core' phage proteins that are widely spread in phages.

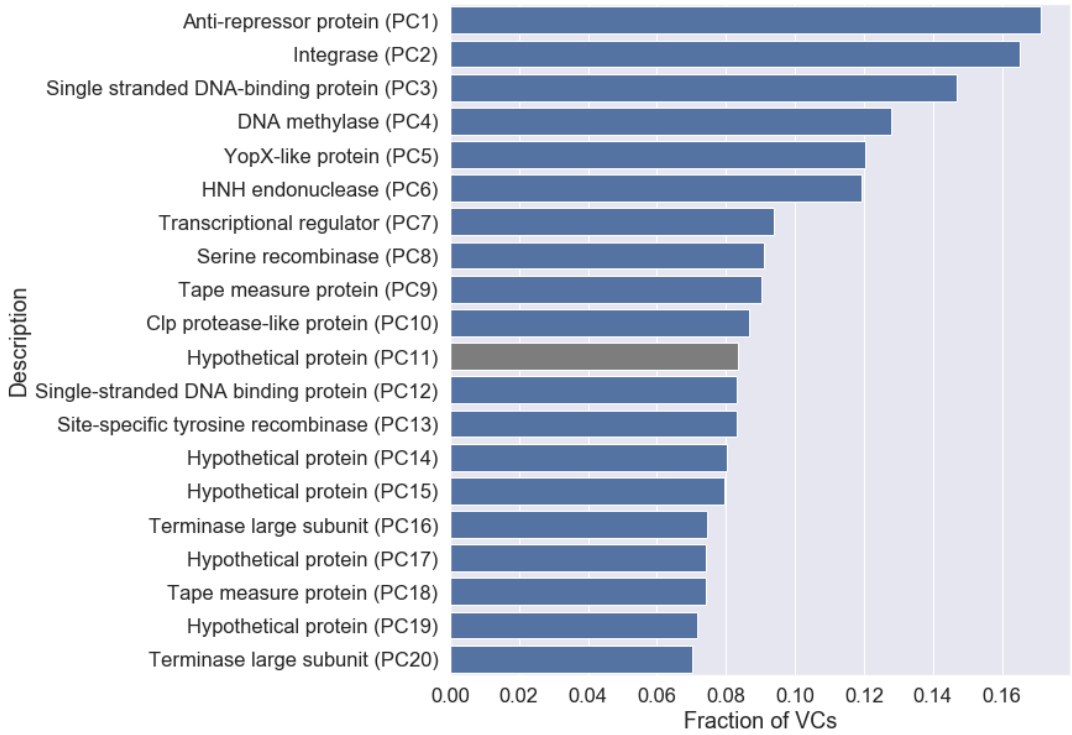


Figure 4.2. Protein clusters (PCs) encoded by gut phages. Prediction of the whole proteome found in GPD and RefSeq phages resulted in the generation of 172,449 PCs. After ranking the PCs by fraction of VCs they were encoded in, the top hits corresponded viral functions such as anti-repressor proteins, integrases, and structural proteins. Interestingly, one of the PCs found in ~8% of the VCs could not be assigned a function based on RefSeq proteins.

4.2.3 Identification of hypervariation domains uncovers putative phage tropism determinants

Prediction of the gene that confers bacterial host specificity to a phage (receptor binding protein) is important for characterization purposes but also because it can be mutagenized to expand the host range (Dunne et al., 2019). The latter is particularly interesting as viruses with broad host range can be harnessed to improve the effectiveness of phage therapy against antibiotic resistant bacteria (Yehl et al., 2019). Receptor binding proteins (RBPs) recognize a bacterial membrane protein (phage receptor) which facilitates adsorption of the phage onto their host (Dowah and Clokie, 2018). As a countermeasure to avoid infection, bacteria often mutate their receptor. However, phages respond by evolving their RBPs to recognize the new receptor. This predator-prey dynamics give rise to hypervariation in the binding domain of the RBPs and the bacterial receptor (Hampton et al., 2020).

I exploited the genetic variation present in the top VC of GPD to identify a candidate RBP for p-crAssphage (Figure 4.3A). After clustering the whole proteome of the crAssphage VC at >70% sequence identity and >90% coverage of both sequences, I sought to quantify amino acid diversity along a cluster of homologous crAssphage proteins. A sudden surge in diversity (hypervariation) would indicate the presence of a binding domain involved in host recognition. I identified such pattern in a group of homologous proteins predicted to be tail fibres. Attachment of tailed phages to bacteria is often mediated by tail fibres and surface receptors, providing further evidence that this set of proteins represent the RBP of p-crAssphage. The spike of amino acid diversity spanned ~70 amino acids and was located at the C-terminus. This finding is consistent with other phage receptor binding proteins that have their hypervariable domains at the C-terminus (Dunne et al., 2019).

I repeated the same exercise but with genomes found in the VC which corresponds to the Gubaphage clade (Figure 4.3B). I identified a large protein (> 2000 amino acids) with a hypervariable region of ~150 amino acids. Proximal genes to this protein included the major capsid protein and the terminase which due to phage modularity tend to be close to tail genes, so the identified protein with an hypervariable domain from Gubaphage is well suited to be a candidate receptor binding protein.

Thus, identification of hypervariable regions can help narrow down the function of important phage genes such as their receptor binding proteins. Elucidation of alternative strategies to homology search can prove invaluable in the characterization of the large fraction of hypothetical proteins in phages.

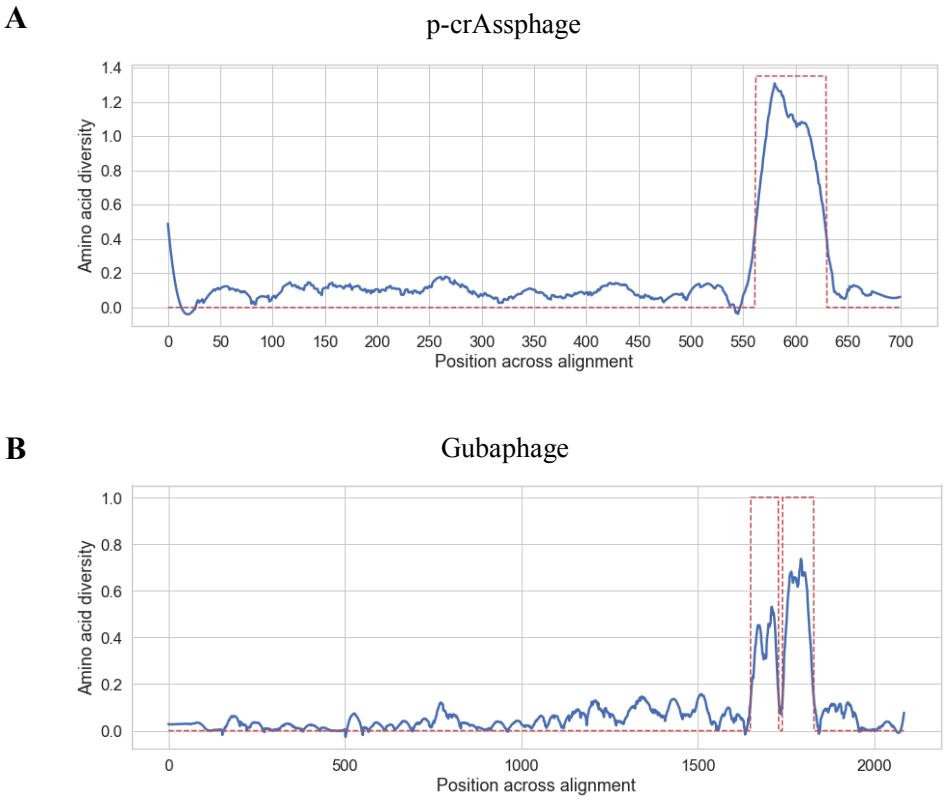


Figure 4.3. Hypervariable domains can narrow down protein function in phages. Detection of hypervariation protein domains can be useful to narrow down protein function in phages. Using this strategy I was able to identify candidate proteins to be the receptor binding proteins of the p-crAssphage **A)** and the Gubaphage clade **B)**.

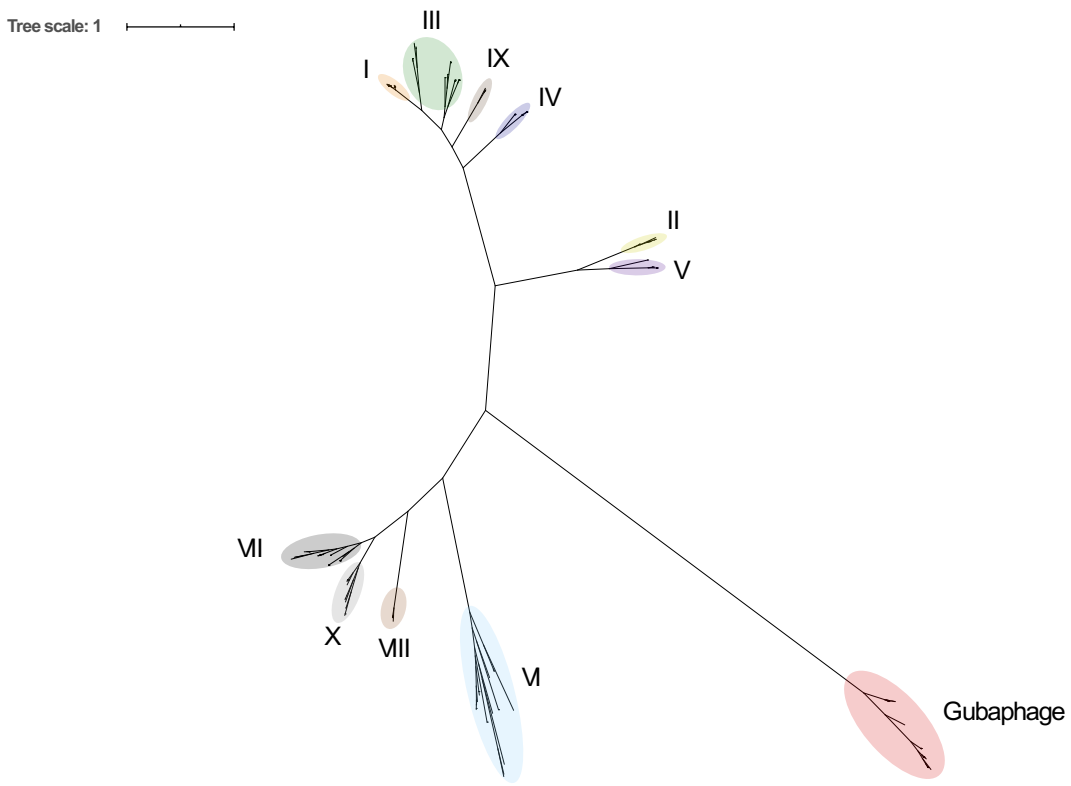
4.2.4 The Gubaphage represents a novel clade of gut phages

As mentioned in the previous chapter, the top two VCs of GPD predictions (p-crAssphage and Gubaphage) represented outliers regarding genetic diversity (as number of genomes / VC). Nucleotide sequence alignment with p-crAssphage revealed no significant similarity. However, they shared some functional features such as large genome size (>80 kb), a BACON domain-containing protein, predicted *Bacteroides* host range, and circular genomes. Searching

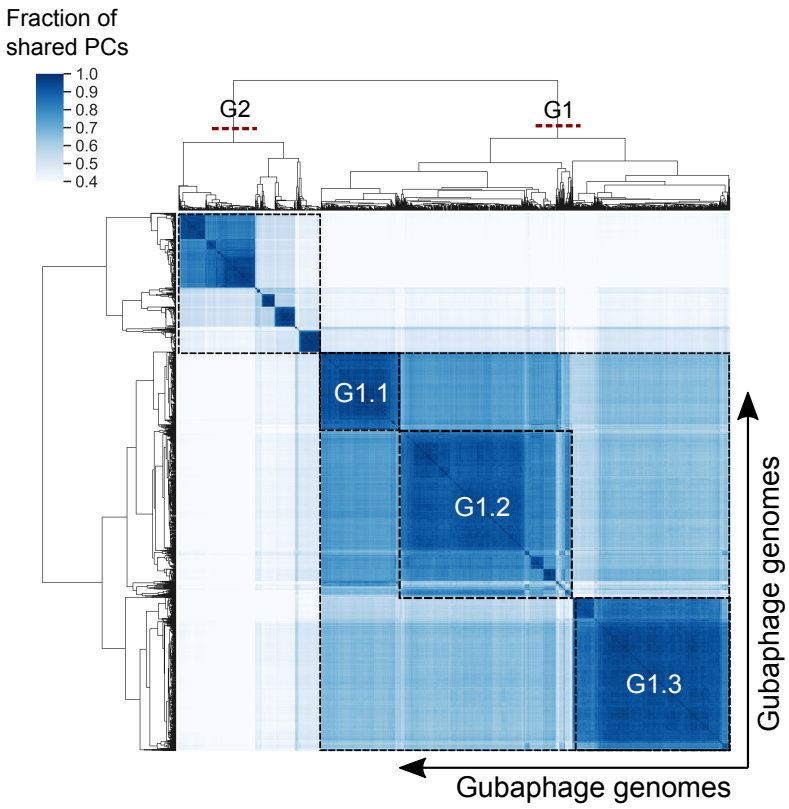
for sequences in the GPD with significant similarity to the Gubaphage large terminase gene (E-value < $1x10^{-6}$), I identified other 205 related VCs. Given its reminiscent features to crAssphage, I decided to investigate if the Gubaphage belonged to the recently proposed crAss-like family which consists of 10 genera and 4 subfamilies (Guerin et al., 2018). I examined this relationship by building a phylogenetic tree using the large terminase gene (Figure 4.4A). The tree successfully clustered all the crAss-like genera as expected, however the Gubaphage significantly diverged from the other crAss-like phages forming a distinct clade.

I then sought to characterize the phylogenetic structure of Gubaphage (Figure 4.4B). Analysis of protein overlap between Gubaphage's genomes revealed that this clade is composed of 2 clusters that share more than 20% but less than 40% of homologous proteins between them. This structure suggests two genera (G1 and G2) from a single viral subfamily. In addition, within G1 I identified another phylogenetic substructure composed of 3 large clusters (G1.1, G1.2, and G1.3) composed of 313, 514, and 502 phage genomes respectively. Host range prediction revealed that G1.1 infects *Bacteroides caccae* and *Bacteroides xylanisolvens* B, G1.3 *Bacteroides* B *vulgatus*, and G2 *Parabacteroides merdae* and *Parabacteroides distasonis*. In the case of G1.2, I couldn't confidently predict a putative host. Interestingly, the larger genetic distance between G1 and G2 also resulted in a more extreme host range switch, from Bacteroidaceae (G1) to Porphyromonadaceae (G2). Core genes of the Gubaphage included homing endonucleases, DNA polymerase I, FluMu terminase, DNA primase, DNA helicase, Thymidylate kinase, dUTPase, among others. Annotation of its genome revealed that Gubaphage is organized into three distinct regions (Figure 4.4C). One region encodes DNA machinery, the second is composed mainly structural genes and the third codes for a series of hypothetical proteins.

A


Tree scale: 1
III
I
IX
IV
II
V
VII
X
VIII
VI
Gubaphage

B


Fraction of shared PCs
1.0
0.9
0.8
0.7
0.6
0.5
0.4
G2
G1
G1.1
G1.2
G1.3
Gubaphage genomes
Gubaphage genomes

C

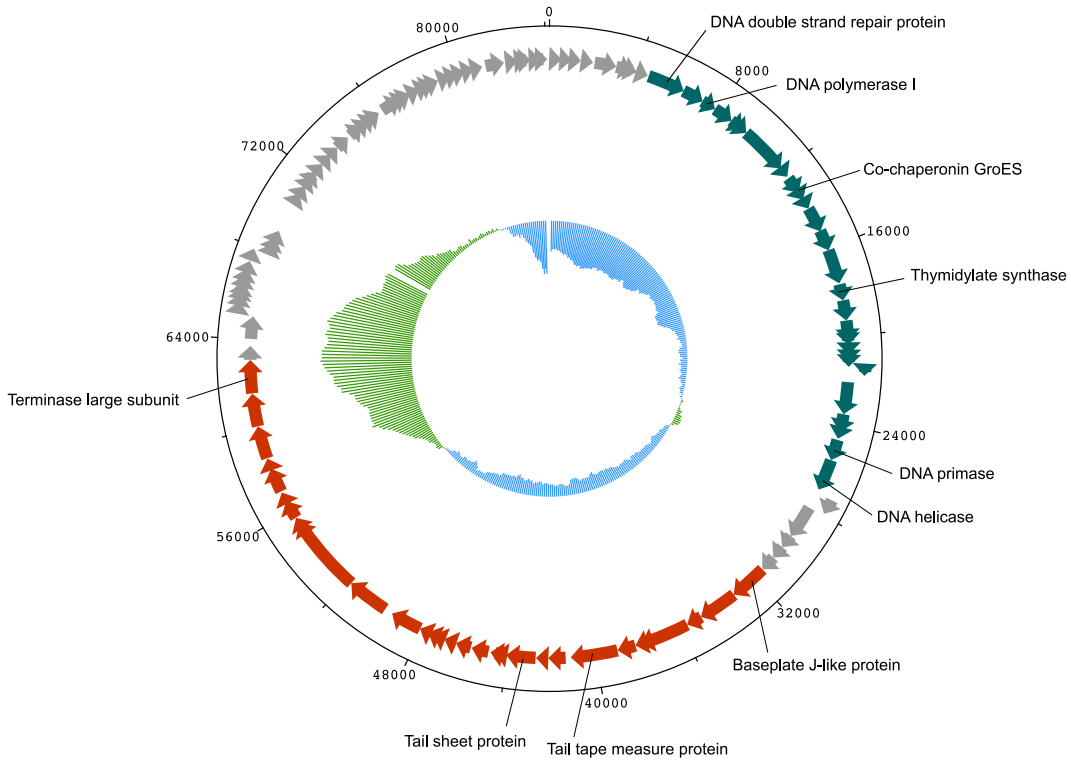


Figure 4.4. The Gubaphage clade. A) Unrooted tree showing the relationship of the crass-like phages and the Gubaphage. Each of the crAss-like clades (I to X), represents a different genus. The Gubaphage forms a clade of its own, suggesting a distant relationship to the crAss-like phages. The tree was constructed by carrying out a multiple alignment of the large terminase genes. **B)** Analysis of Gubaphage phylogenetic structure revealed two genera infecting member of the *Bacteroides* (G1) and *Parabacteroides* (G2) genera. **C)** Inspection of Gubaphage genome reveals that it is composed of 3 parts. The first one (blue-green) codes for DNA machinery, the second (red) harbours structural proteins such as the large terminase, and tail proteins, the third (grey top left) consists of only hypothetical proteins. Inner bars represents GC skew.

4.2.5 Expansion of the *Picovirinae* subfamily

Hitherto I have focused on novel phage clades (crAss-like family and Gubaphage clade), however phages belonging to traditional phage subfamilies such as *Spounavirinae*, *Peduovirinae*, *Autographivirinae*, and *Picovirinae* have been detected in human faces (Waller

et al., 2014). I decided to explore the diversity of the *Picovirinae* subfamily because it was one of the most common taxa predicted in GPD.

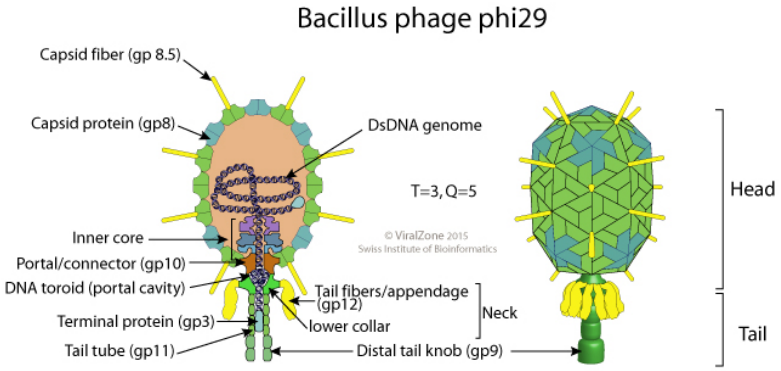
Picovirinae phages are known to have a small linear double stranded DNA genome of about 16-20 kb. They belong to the *Caudovirales* order and have an icosahedral capsid with a non-contractile tail (Figure 4.5A). The *Picovirinae* subfamily is currently composed of 3 genera namely *Salasvirus*, *Negarvirus*, and *Cepanuvirus* (Hulo et al., 2011). I predicted all the phages in GPD from this family by using a marker gene approach and obtained 4807 genomes.

In order to study the phylogenetic structure of the recovered genomes, I calculated all the pairwise overlaps of protein clusters between the *Picovirinae* genomes. Interestingly, after clustering the genomes and visualizing them in a heatmap, a phylogenetic substructure consisting of 4 large clades emerged (Figure 4.5B). Furthermore, an unrooted tree inferred from the PCs overlap clearly suggested 4 clades (Figure 4.5C). Given this evidence, I decided to structure the *Picovirinae* subfamily into 4 clades: Picovirinae_1 (P1), Picovirinae_2 (P2), Picovirinae_3 (P3), and Picovirinae_4 (P4). In addition, P1 clade was clearly divided into two clades, Picovirinae_1_1 (P1_1) and Picovirinae_1_2 (P1_2). With this new structure I was able to assign a clade to the three classified genera, while *Salasvirus* were assigned to P2, *Cepanuvirus* and *Negarvirus* were assigned to P1_1. In addition, I assigned a clade to several unclassified members of the *Picovirinae* with this expanded phylogenetic structure. Notably, P1_2, P3, and P4 remained without any known *Picovirinae* phage members assigned to them.

Host assignment revealed more than 288 gut bacteria isolates distributed between the Firmicutes and Actinobacteriota, moreover, P1_2, P3 and P4 were restricted to the Firmicutes, leaving P1_1 as the only inter-phyla *Picovirinae* clade. Containment of phage clades to a specific phylum is expected, as very distantly related host bacteria can present challenges to polyvalent phages e.g. substantially different replication machinery. In total, 31 genera of the human gut microbiota were predicted to be susceptible to infection by *Picovirinae* phages (Figure 4.5D).

This finding represents a clear example of the importance of metagenomics to fill in viral diversity gaps. In addition, gaining further knowledge of *Picovirinae* phages is important because their lytic lifestyle is suitable for phage therapy directed to Actinobacteriota and the Firmicutes.

A



B

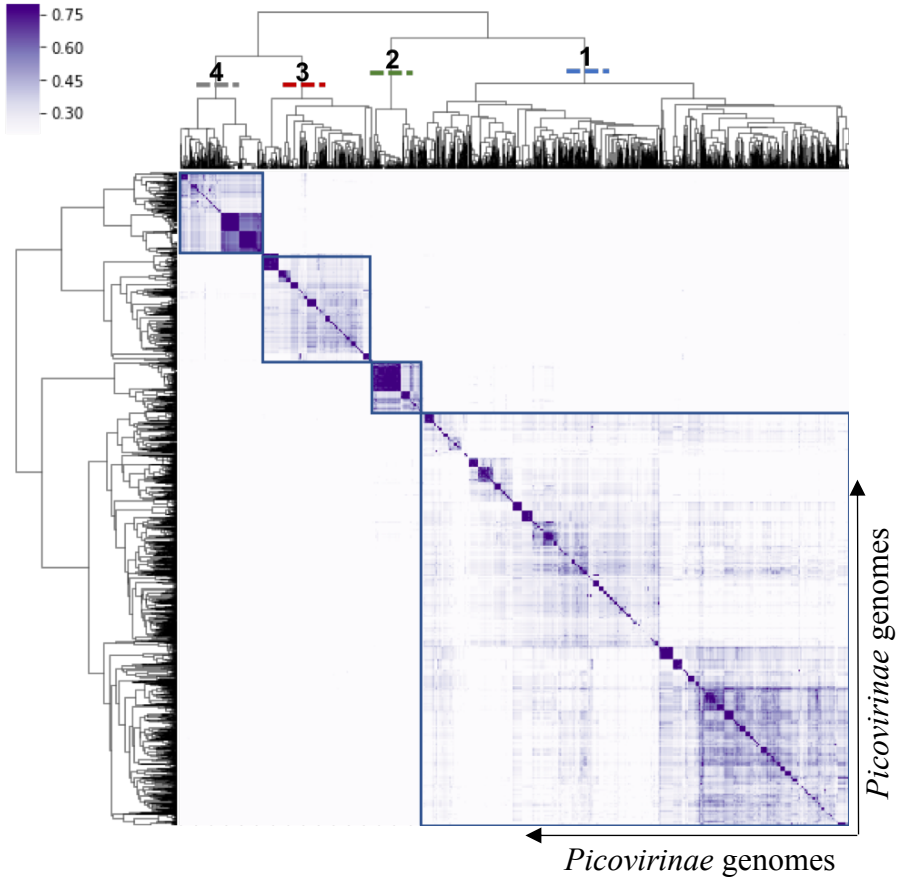
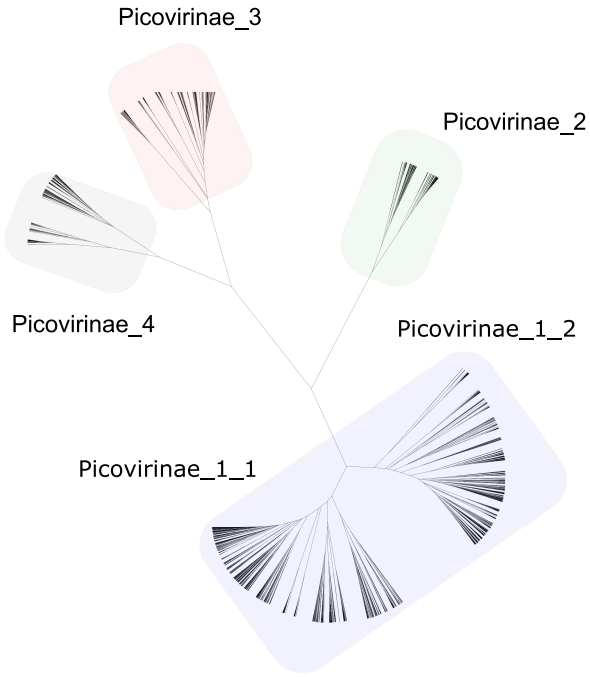


Figure 4.5. Expansion of the *Picovirinae* subfamily. A) The *Picovirinae* subfamily is characterized by having relatively small genomes (16-20kb) and a lytic lifecycle. They possess a linear double stranded DNA and have an icosahedral capsid with a non-contractile tail. **B)** Analysis of the phylogenetic structure of gut *Picovirinae* phages by fraction of shared protein clusters suggested 4 large clades.

C



D

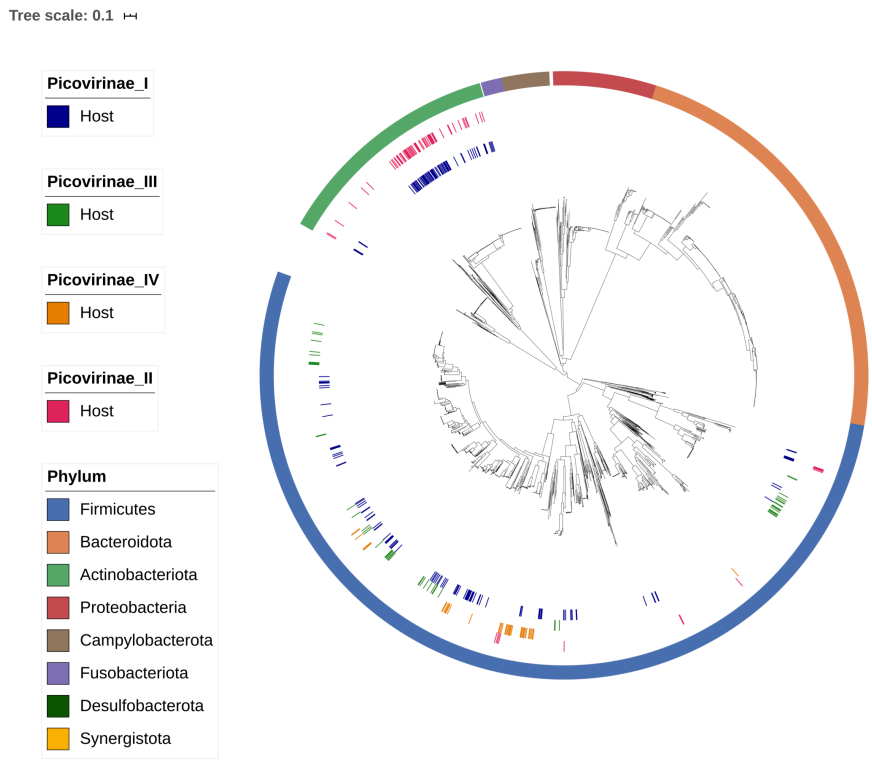


Figure 4.5. Expansion of the *Picovirinae* subfamily. C) Unrooted tree of shared protein clusters. The 4 clades were named Picovirinae_1, Picovirinae_2, Picovirinae_3, Picovirinae_4. This expanded diversity of the *Picovirinae* was able to accommodate the 3 known genera and several unclassified phages. Notably, Picovirinae 3 and 4 represented completely novel clades.

The tree was generated by calculating the fraction of shared protein clusters among individual Picovirinae phages and then carrying out hierarchical clustering with average linkage and Euclidean metric. **D)** Host assignment of *Picovirinae* phages to gut bacteria. Hosts were predicted by CRISPR spacer exact matching and prophage assignment. The tree was built by concatenating 40 universal core marker genes from each of the 2898 gut bacteria isolates and then carrying out a multiple sequence alignment. P1_2, P3 and P4 were restricted to the Firmicutes, leaving P1_1 as the only inter-phyla *Picovirinae* clade (Firmicutes and Actinobacteriota host range).

4.2.6 Viral diversity across gut bacteria clades

I next inferred the most likely bacterial hosts for each phage prediction using a comprehensive collection of 2898 human gut microbiota isolate genomes. By screening for the presence of CRISPR spacers (Edwards et al., 2016) targeting phage and by linking the prophages to their assemblies of origin, I was able to carry out host assignment. In order to estimate the rate of false positives (FPs) due to CRISPR random matches, I generated synthetic random spacers and mapped them against the GPD. Repeating this procedure 100 times revealed the distribution of the expected number of FPs across different matching criteria (Figure 4.6A). As can be seen from the graphs, no FPs are detected due to random chance when no mismatches are allowed across the whole length of the spacer (the criteria used in this work for the original mapping). However, as more mismatches are allowed, there is an increase in random matches across all coverages tested. Notably, at 80% coverage and only 4 mismatches allowed, the expected false positive rate due to random chance reach 2.6% of all the matches reported from the original mapping.

In total, I assigned 2,157 hosts to 40,932 GPD phage (28.66% of all predictions). This corresponded to at least one phage for 74.43% of all cultured human gut bacteria. I then analysed if there was any preference for phage infection across 5 common human gut bacterial phyla (Firmicutes, Bacteroides, Proteobacteria, and Actinobacteriota). At the phylum level, I detected significant lower phage prevalence in Actinobacteriota, with 58.79% infected isolates compared to at least 70% for the other phyla (Figure 4.6B).

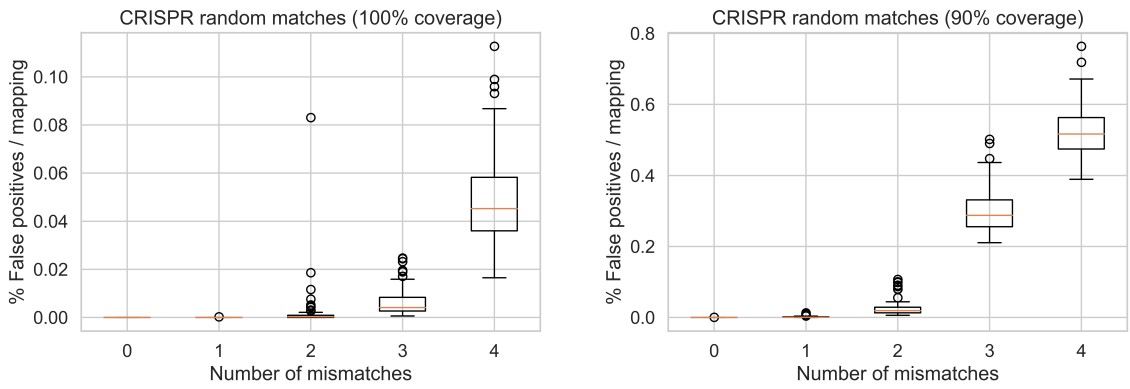
I then measured viral diversity (measured by the number of VCs per isolate) within each phylum (Figure 4.6C). This analysis revealed that the Firmicutes harbour a significantly higher

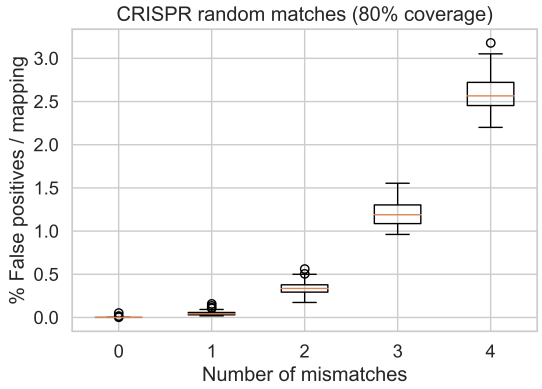
viral diversity, with an average of 3.13 VCs/isolate while also harbouring 60% of the total VCs assigned across all phyla. Interestingly, the Firmicutes diversity was unevenly distributed as most of the viral diversity originated from the Negativicutes and Clostridia classes, with an average of 4.88 VCs and 3.9 VCs per isolate in contrast with the Bacilli (0.99 VCs/isolate), and none for *Bacilli*_A and Desulfitobacteriia classes.

Analysis at the bacterial genus level across all phyla revealed that *Lachnospira*, *Roseburia*, *Agathobacte*r, *Prevotella*, and *Blautia*_A host the highest number of VCs/isolate (Figure 4.6D). With the exception of *Prevotella*, which belongs to the Gram-negative Prevotellaceae family, these genera are members of the Gram-positive Lachnospiraceae family of Firmicutes associated with butyrate-producing spore-formers. In contrast, the lowest viral diversity per isolate was detected among *Helicobacter*, and the lactic acid bacteria *Lactobacillus*, *Lactobacillus*_H, *Enterococcus*_D and *Pediococcus.* Thus, I observe a wide distribution of phage abundance and prevalence across human gut bacteria, even within the same phylum.

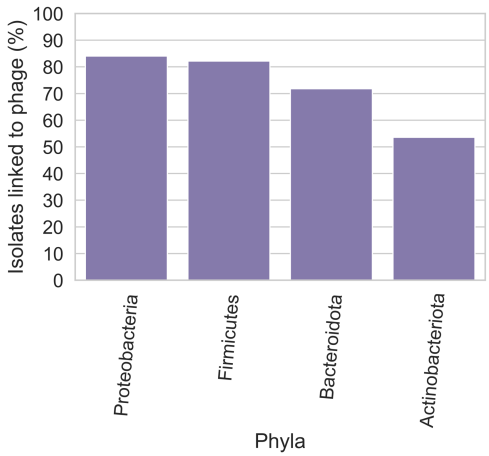
CRISPR spacers can be used to link phages with their host but a limitation is that some bacteria do not encode them and thus their phages will not be detected in the analysis. Although it's estimated that around 46% of bacteria code for CRISPR systems (Karginov and Hannon, 2010), I detected CRISPR spacers in 56.36% of the gut isolate genomes. Despite the discrepancy with the previous estimate, a larger prevalence in the gut may be plausible. It's possible that the incidence of CRISPR systems may vary across different environmental niches.

A

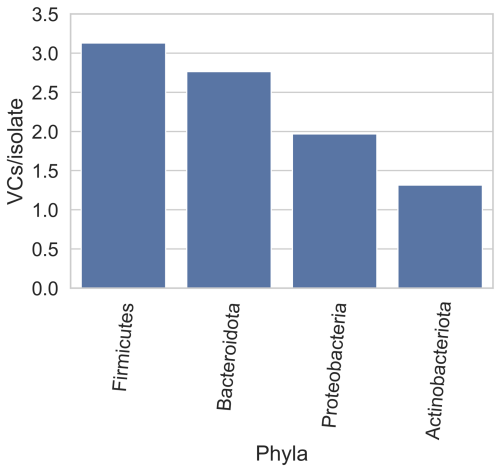



CRISPR random matches (80% coverage)
% False positives / mapping
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
Number of mismatches

B


Isolates linked to phage (%)
0
10
20
30
40
50
60
70
80
90
100
Proteobacteria
Firmicutes
Bacteroidota
Actinobacteriota
Phyla

C


VCs/isolate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Firmicutes
Bacteroidota
Proteobacteria
Actinobacteriota
Phyla

D

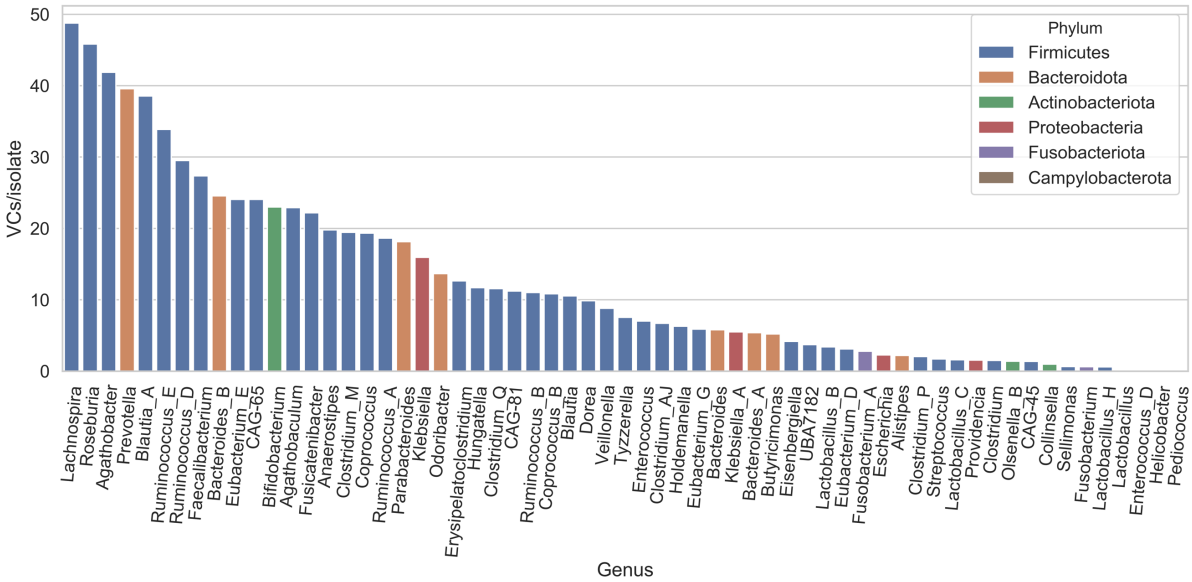

VCs/isolate
0
10
20
30
40
50
Phylum
Firmicutes
Bacteroidota
Actinobacteriota
Proteobacteria
Fusobacteriota
Campylobacterota
Lachnospira
Roseburia
Agathobacter
Prevotella
Blautia_A
Ruminococcus_E
Ruminococcus_D
Faecalibacterium
Bacteroides_B
Eubacterium_E
CAG-65
Bifidobacterium
Agathobaculum
Fusicatenibacter
Anaerostipes
Clostridium_M
Coprococcus
Ruminococcus_A
Parabacteroides
Klebsiella
Odoribacter
Erysipelatoclostridium
Hungatella
Clostridium_Q
CAG-81
Ruminococcus_B
Coprococcus_B
Blautia
Dorea
Veillonella
Tyzzerella
Enterococcus
Clostridium_AJ
Holdemanella
Eubacterium_G
Bacteroides
Klebsiella_A
Bacteroides_A
Butyricimonas
Eisenbergiella
UBA7182
Lactobacillus_B
Eubacterium_D
Fusobacterium_A
Escherichia
Alistipes
Clostridium_P
Streptococcus
Lactobacillus_C
Providencia
Clostridium
Olsenella_B
CAG-45
Collinsella
Sellimonas
Fusobacterium
Lactobacillus_H
Lactobacillus
Enterococcus_D
Helicobacter
Pediococcus
Genus

Figure 4.6. Viral diversity across gut bacteria clades. A) In order to quantify the rate of FPs due to CRISPR random matches, I generated 100 sets of synthetic random spacers and mapped them against the GPD. No FPs were detected at 100% coverage and no mismatches allowed. Across all coverages tested, the rate of FPs increased as more mismatches were allowed. **B)** Percentage of isolates of each phylum linked to phage. Actinobacteriota had the lowest percentage of isolates predicted to be a phage host. Actinobacteriota vs Bacteroidota (P = 0.007, χ^2 test), Actinobacteriota vs Proteobacteria (P = 0.0025, χ^2 test), Actinobacteriota vs Firmicutes (P = 1.01 x 10^{-5}, χ^2 test). **C)** The Firmicutes hosted the highest viral diversity (highest number of VCs/isolate). Firmicutes vs Bacteroidota (P = 0.021, χ^2test), Firmicutes vs Proteobacteria (P = 4.41 x 10^{-6}, χ^2 test), Firmicutes vs Actinobacteriota (P = 1.1 x 10^{-31}, χ^2test). **D)** Bacterial genera with the highest viral diversity were *Lachnospira*, *Roseburia*, *Agathobacter*, *Prevotella*, and *Blautia*_A. On the other hand, the lowest viral diversity was harboured by *Helicobacter* and the lactic acid bacteria *Lactobacillus*, *Lactobacillus*_H, *Enterococcus*_D and *Pediococcus*.

4.2.7 Evaluating host range of gut phages

Horizontal transfer of genes between bacteria via transduction is a major driver of gene flow in bacterial communities (Chen et al., 2018). Host tropism of bacteriophage is believed to be limited by phylogenetic barriers, with most phages being usually restricted to a single host bacterial species (Ackermann, 1998). However, this has not been investigated at large scale across the human gut bacteria. Host assignment at different bacterial taxonomic ranks revealed that the majority of VCs were restricted to infect a single species (64.51%) (Figure 4.7A). I also found many VCs with broader host ranges such as those restricted to a single genus (22.39%), family (10.79%), order (1.86%), class (0.26%) and phylum (0.13%). These findings are in line with a recent survey of the host range of gut phages by meta3C proximity ligation (6,651 unique host-phage pairs) which found that ~69% of gut phages were restricted to a single species (Marbouty et al., 2020). Visualization of very broad range VCs (i.e. those not restricted to a single genus) reveals the large-scale connectivity between phylogenetically distinct bacterial species (Figure 4.7B).

In general, the higher the viral diversity per bacterial genus, the higher the number of phages with broad host range (Spearman's Rho = 0.6685, P= 3.91x10^{-9}) (Figure 4.7C). Even though

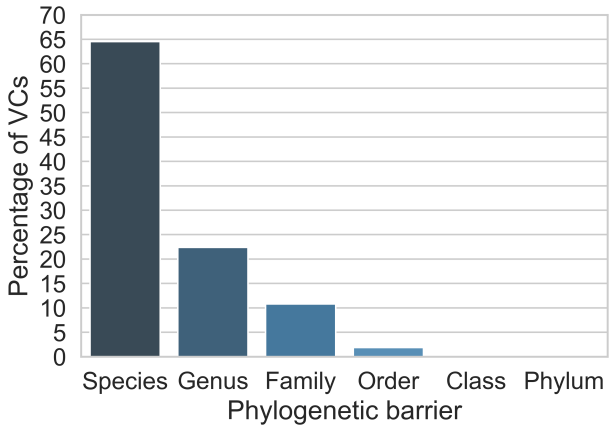
this trend could be explained due to the presence of random matches, as discussed above, no FPs were detected using perfect matches. In addition, when I permuted the labels of the host assignment 300 times, I found the original linear model to significantly deviate from the random one ($P < 0.001$). The average number of broad host range hits for the permuted assignments was 726.9 versus 38.344 for the original assignment, highlighting the containment of phages within bacterial clades.

Surprisingly, two VCs (VC_269 and VC_644) had a host range that spanned two bacterial phyla. VC_269 was predicted to infect *Faecalibacterium prausnitzii*_C (Firmicutes) and two *Bifidobacterium spp.* (Actinobacteriota), while VC_644 had a host range that included 5 *Bacteroides spp*. (Bacteroidota) and *Blautia*_A *wexlerae* (Firmicutes). I predicted VC_269 to be a *Myoviridae* phage, on the other hand, I could not assign a taxonomy rank to VC_644. The presence of integrases in both VCs suggest that these are temperate phages. I hypothesize that additional phages infecting both Actinobacteriota and Firmicutes may be more common, as recent evidence supports a shared ancestry between phages that infect both Actinobacteriota (*Streptomyces*) and Firmicutes (*Faecalibacterium*) (Koert et al., 2019).

Taken together, I reveal that approximately one third of gut phage have a broad host range not limited to a single host species. This analysis provides a comprehensive blueprint of potential phage mediated gene flow networks in human gut microbiome.

The emergence of broad host range phages or 'generalists' has been linked with shifts in bacterial composition linked to nutrient availability (Warwick-Dugdale et al., 2019). In addition, phage generalism has been associated with lower infection efficiency (Howard-Varona et al., 2018). Many members of the gut microbiome are considered copiotrophs based on the copy number of the Ribosomal RNA operon (rrn), as it positively correlates with cellular ribosomal content and maximum growth rate (Gao and Wu, 2018). This would imply that in general, the gut is not a limited nutrient environment and phages can 'secure' a stable host. As stated, the majority of the viral diversity reported here was predicted to infect a single species, which is in line with copiotroph hosts. It's important to consider that some gut bacteria may be oligotrophs as it's increasingly recognized that nutrients in the gut vary spatially (Donaldson et al., 2016). This scenario would probably result in a higher proportion of broad host range for some bacterial species.

A



B

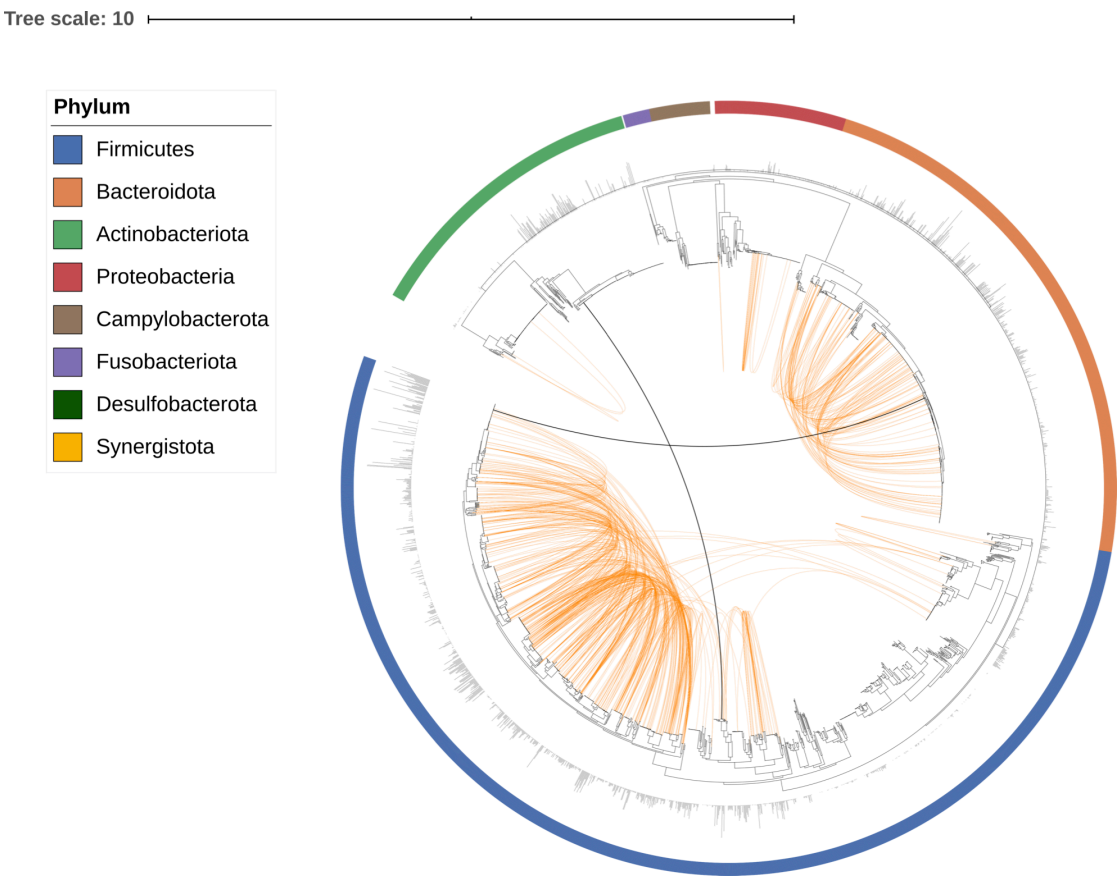


Figure 4.7. Host range of gut phages. A) The majority of VCs were found to be restricted to infect a single species ($P = 0.0$,binomial test). However, a considerable number of VCs (~36%) had a broader host range. **B)** Phylogenetic tree of 2898 gut bacteria isolates showing phage host range. Host assignment was carried out by linking prophages with their assemblies and CRISPR spacer matching. Orange connections represent VCs not restricted to a single genus). Black connections represent VCs able to infect two phyla. Outer bars show phage diversity (VCs/isolate).

C

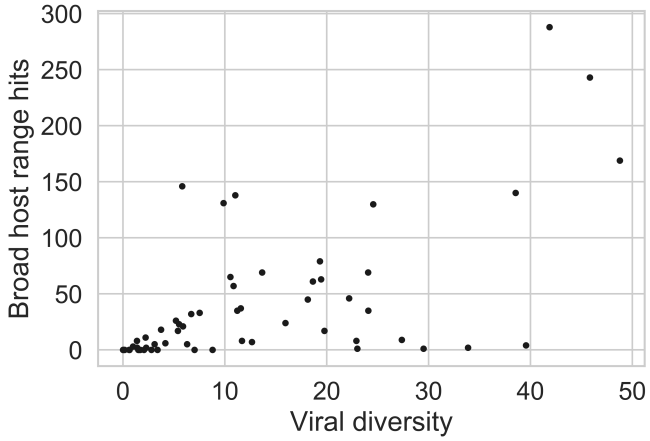


Figure 4.7. Host range of gut phages. C) In general, the higher the viral diversity per bacterial genus, the higher the number of phages with broad host range (Spearman's Rho = 0.6685, P= 3.91x10-9). This trend was significantly different than the one generated from permuting the host assignment labels ($P < 0.001$).

4.3 Conclusions

In this chapter, I carried out a large-scale analysis of gut phages to shed light into their encoded functions. Top viral functions were primarily involved in basic functions of the life cycle of phages such as replication, virion assembly, and lytic enzymes. However, a particular interest of mine was to explore the possibility of gut phages carrying non-canonical viral proteins. In that regard, I found several clades of phages encoding enzymes that participate in sulphur and nucleotide metabolism.

I expect that many of these non-classical viral proteins are involved in promoting a successful infection by energy generation (dissimilatory sulfate reduction) or by manipulating the bacterial nucleotide pool to avoid misincorporation of uracil into the genome of DNA phages. I found that gut phages commonly encode reverse transcriptases (RTs) (~13% of VCs) as opposed to RefSeq phages (<1%). These viral RTs may be fulfilling critical roles in gut phages such as generation of sequence diversity in their receptor binding proteins (RBPs) and protecting lysogens from infection by other phages (superinfection immunity). I also discovered other rare instances (<0.5% of VCs) of phages encoding nutrient uptake genes (e.g. taurine, zinc) which may be of benefit to the bacterial host.

A common issue when analysing metagenomics data is the significant number of proteins annotated as 'hypothetical', hindering efforts to carry out comprehensive functional analyses. This problem is further exacerbated with phages, in part due their large genetic diversity and because many functional experiments have been carried out only in a handful of bacteriophage models (e.g, T4, T7, λ phage). For instance, I found a family of hypothetical proteins present in ~8.5% of all VCs. This observation reflected the lack of annotation for even widespread phage proteins. Despite the limitation regarding functional annotation, I explored the possibility of predicting function for hypothetical viral proteins by exploiting hypervariation motifs. This analysis is particularly suitable for the prediction of RBPs in phages given that the binding domain of RBPs is often under selection to overcome mutations in the bacterial receptor. Using this strategy I was able to identify RBP candidates for two of the most genetically diverse phages in GPD (as measured by genomes per VC), namely the p-crAssphage and the Gubaphage. As hypervariation domains are often found in phages, this

analysis provides a powerful way to narrow down gene function in phages when there is enough availability of viral genetic diversity.

In this chapter I also analysed the Gubaphage clade in detail. Despite the lack of sequence similarity of Gubaphage to p-crAssphage, these phages shared other functional features such as large genome size (>80 kb), *Bacteroides* host range, a BACON-containing protein and a circular genome. Given the high variation of the crAss-like family, these features prompted me to investigate if Gubaphage belonged to a current or novel crAss-like genus or if it was a completely novel clade. By compiling a list of genomes representing all the crAssphage genetic diversity and then constructing a tree using terminase large subunit gene, I discovered that the Gubaphage did not fit any of the previous crAssphage clades. Another interesting feature of Gubaphage was the high number of genomes associated to its VC, suggesting its high prevalence in human metagenomes. Indeed, in the next chapter I use more sensitive methods to confirm its high prevalence across human populations. Elucidation of the functional traits of Gubaphage will require its isolation and characterization as this will help to establish a clearer view of its role in the human gut microbiome.

Having investigated a novel clade of gut phages, I decided to explore the possibility of expanding the diversity of a known phage clade, namely the *Picovirinae* subfamily. In order to study the phylogenetic structure of *Picovirinae* gut phages I computed the fraction of shared PCs among them. This analysis uncovered 4 major phage clades. Notably, all RefSeq classified and several unclassified *Picovirinae* phages were assigned to one of the 4 clades. However, two major clades remained composed of only phages found in GPD. The expansion in diversity of the *Picovirinae* subfamily showcases the importance of metagenomics in filling in diversity gaps in phage taxonomy.

Given the technical challenges when culturing gut bacteria, host assignment of gut phages remains largely unexplored. I opted for two strategies namely CRISPR and prophage matching and in order to minimize false positives, I only considered exact matching. This analysis allowed me to explore viral diversity patterns across different bacterial taxonomic groups. For instance, I found that viral diversity was highest in the Firmicutes while at the genus level, *Lachnospira*, *Roseburia*, and *Agathobacter* harboured the highest number of VCs/isolate, whereas *Enterococcus*_D*, Helicobacter* and *Pediococcus* the least. Notably, I considerably increased the number of phages assigned to less studied bacterial clades. For instance, a search

on “NCBI virus” of phages infecting *Lachnospiraceae* bacteria returns only 8 hits. On the other hand, on this thesis I predicted 2,985 VCs that infected *Lachnospiraceae* bacteria (with an estimated median phage genome completeness of 81.62%).

Although the majority of VCs were found to be restricted to a single bacterial species, a significant percentage (~36%) was predicted to infect multiple species, genera, families, orders, and even classes. A consequence of broad host range phages is an increased connectivity for horizontal gene transfer events between gut bacteria. Since phages can carry genes from their hosts by transduction, broad host range phages can play critical roles in “gene spillage” across very different bacterial clades from the gut microbiome. For instance, a phage can transduce genes from a different family into another bacterial clade. In another transduction event, narrow host range phages (which are more common), can help to move the newly acquired gene into the clade. These events can have important roles in bacterial adaptation in the human gut.