Chapter 5: Global distribution and epidemiology of gut phages

5.1 Introduction and aims

Much of human microbiome research across populations has focused on gut bacteria. Samples from different countries (mainly Western ones), have been analysed for differences in bacterial composition related to health and disease states. In addition, patterns of bacterial profiles have been linked to different factors such as antibiotic use, urbanization, and age. However, epidemiology research of gut phages has been limited and carried out in small cohorts with narrow geographical distribution of samples. Findings to date, include the association of the gut phageome with health and disease, as well as the suggestion of a set of phages carried by at least half of the human population (core virome) (Manrique et al., 2016).

Regarding individual phage clades, efforts have been mainly directed to the analysis of the abundant crAss-like family. For instance, one of the largest studies that analysed the global distribution of crAssphage strains found strong correlations with different clades of gut bacteria, weak associations with diet, but no significant association with health and disease (Edwards et al., 2019).

In this chapter, I analyse global patterns of the human gut phageome and its association with lifestyle and bacterial composition. I then focus on specific VCs, such as those that are widespread across human populations (global) and those that are highly prevalent in individual continents. Finally, I explore the concept of the controversial idea of a core virome using my dataset.

The aims of the research presented in this chapter are:

- assess global patterns of the human gut phageome;
- analyse geographical distribution of relevant VCs;
- assess the concept of a core virome.

5.2 Results and discussion

5.2.1 Saturation curves for VCs

Before proceeding with the analysis of global gut phageome patterns, it was important to assess how much of total viral diversity was captured by GPD predictions (Figure 5.1). With that end, I calculated the number of novel VCs accumulated with the addition of every new sample. By analysing the growth rate of the resultant curve it's possible to estimate the degree of diversity saturation. At the worldwide scale, it seems that GPD reached saturation regarding novel phage diversity. However, this pattern mostly reflects Western continents (64.2% of the samples). When I stratified by continent, in line with the previous finding, Europe and North America seemed to have plateaued. In addition, Asia's and Oceania's curves also showed signs of diversity saturation. In the case of Africa and South America, the diversity appeared to be growing in a linear fashion with each new additional sample, indicating a low degree of saturation. The latter result was expected as the gut phageome of both continents was estimated from only ~200 samples each as opposed to the other continents with thousands of samples. Thus, GPD captured better phage diversity in North America, Europe, Asia and Oceania, while the gut phageome from generally understudied continents such as Africa and South America still remains to be further explored. Importantly, small phages with a genome size < 10 kb (e.g. *Microviridae*) and RNA phages need to be considered for all continents in order to have a fuller picture of the diversity of the gut phageome.

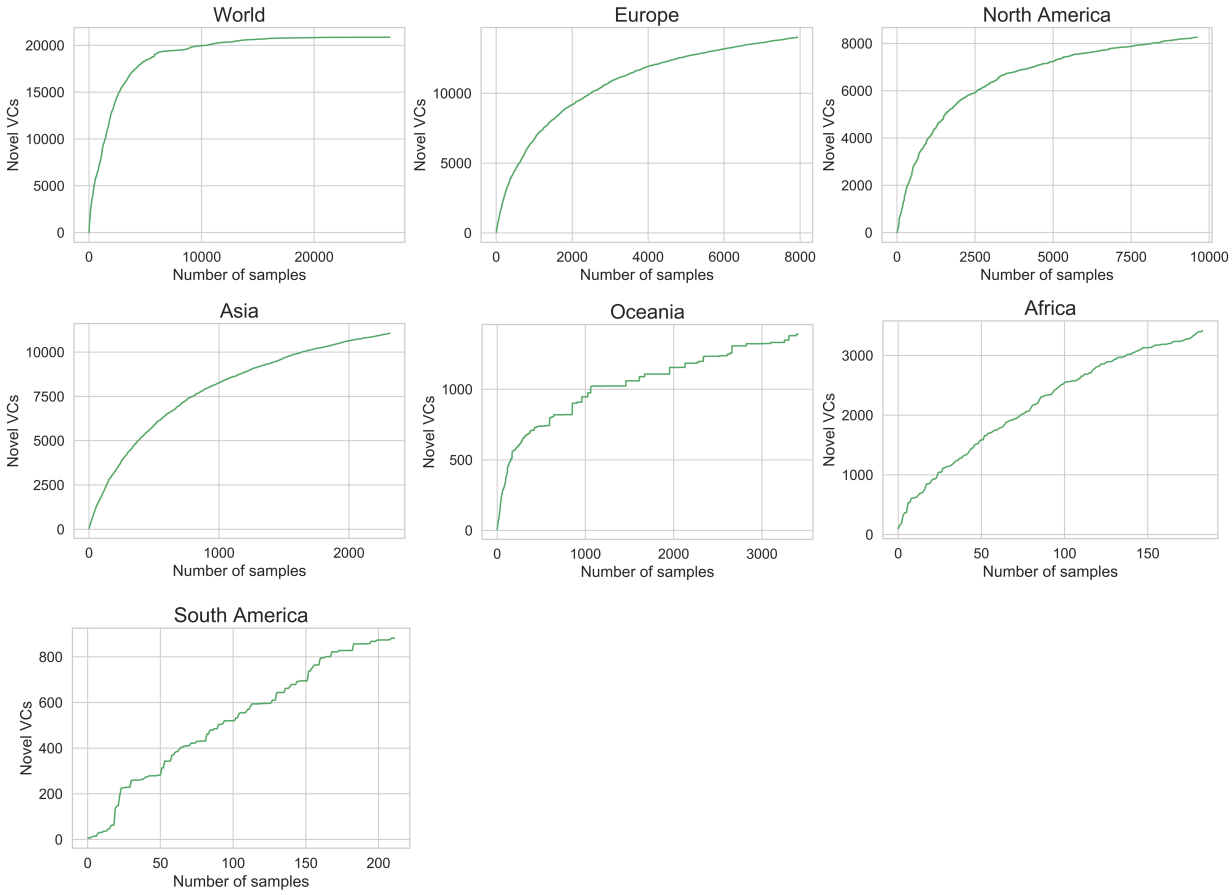


Figure 5.1. Rarefaction curves for viral richness. Saturation curve for viral richness captured in GPD. At the worldwide scale, viral richness seems to have plateaued. However, analysis of individual continents show phage diversity in Africa, and South America is still growing.

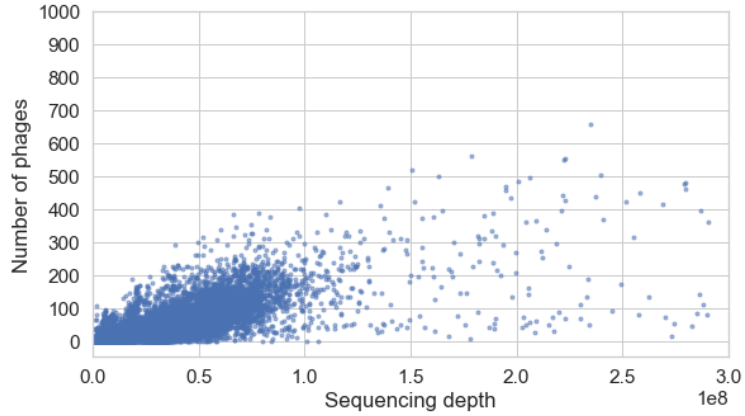
5.2.2 Human lifestyle associated with global gut distribution of phageome types

Each human harbours diverse populations of gut phage, referred to as a phageome. The 28,060 metagenomic datasets used to generate the GPD were sampled from 28 different countries across the six major continents (Africa, Asia, Europe, North America, South America and Oceania) providing a basis to explore patterns in gut phageomes across human populations. I removed samples with a sequencing depth below 50 million reads/sample, as below this threshold I observed a positive correlation between sample depth and number of viral genomes detected (Figure 5.2A). This new subset consisted of 3011 samples and spanned all the continents and 23 countries. I estimated the similarity between samples by computing the number of shared VCs and normalizing it by the total number of VCs in both samples (Jaccard index).

I observed that North American, European, and Asian samples segregated from African and South American samples (Figure 5.2B). Interestingly, this pattern is associated with important differences in human lifestyles. Country-wise, samples derived from Africa and South America come mainly from Peru, Tanzania, and Madagascar. Specifically, Peruvian and Tanzanian samples originate from hunter gatherer communities whereas Malagasy samples come from rural communities with non-Western lifestyles. Oceania was a special case because it had a similar fraction of samples belonging to both groups. However, when I stratified by country, all Fijian samples went to the rural group, whereas Australian samples segregated with the urbanized cluster. Fiji samples were derived from rural agrarian communities. These observations support the hypothesis that lifestyle, particularly urbanization, may drive differences in the gut phageome across different human populations.

I reasoned that the bacterial composition of an individual's microbiome would shape the gut phageome. Prevotellaceae bacteria are more abundant and prevalent in individuals living a rural/traditional lifestyle, whereas *Bacteroides* are more abundant and prevalent in individuals living a urban/Western lifestyle (Wu et al., 2011). By harnessing the host assignment data for each phage, I found that the proportion of VCs assigned to the Prevotellaceae family from African, South American and Fijian samples was much higher than that of North America, Europe, Asia, and Australia (Figure 5.2C). I observed an inverse relationship with *Bacteroides* phage, which were significantly more prevalent in North America, Europe, Asia, and Australia gut microbiomes. Given the correlation of enterotypes and phageome types, driven by the intimate connection between phages and their bacterial hosts, I provide evidence that human lifestyle drives global patterns of gut phageomes by mediating changes in the bacterial gut microbiome.

A



B

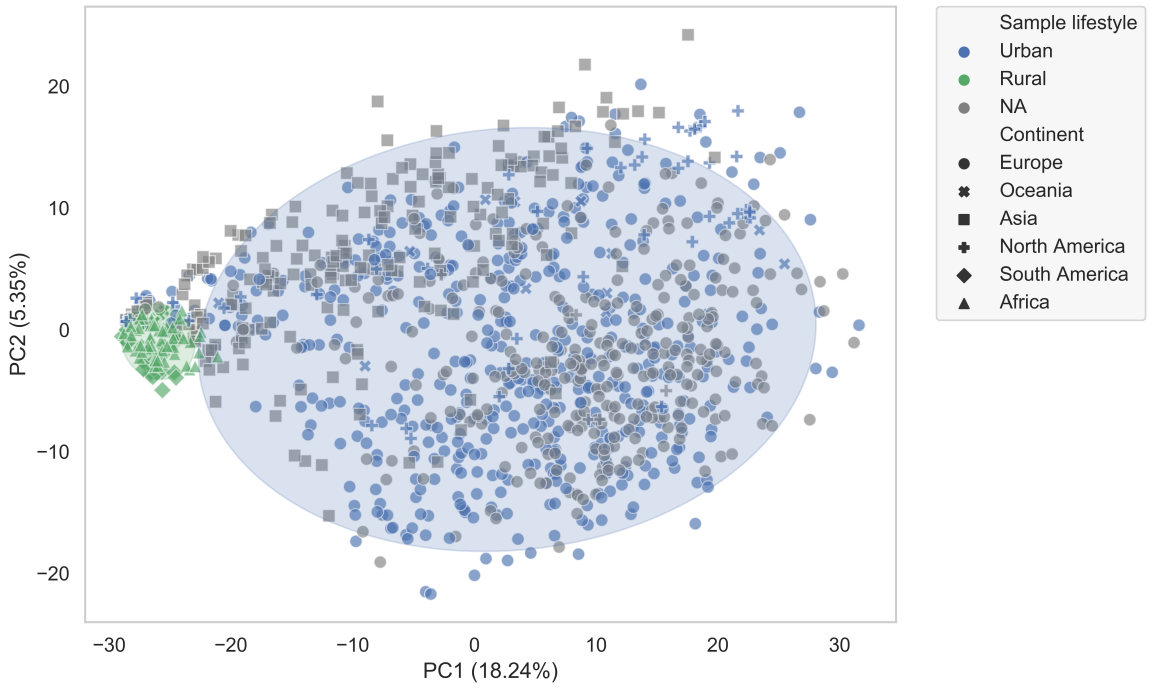


Figure 5.2. Human lifestyle is associated with global gut distribution of phageome types. **A)** Samples exhibit a positive correlation between sequencing depth and number of phage genomes detected. Correlation of samples with sequencing depth < 50 million (Pearson's r: 0.6825, *P* = 0.0). Correlation of samples with sequencing depth > 50 million (Pearson's r: 0.3681, *P* = 2.79e-97). **B)** PCA plot of inter-sample Jaccard distance. Lifestyle is associated with differences in the gut phageome across human populations. Samples from Peru, Madagascar, Tanzania and Fiji are found in the rural cluster whereas those samples with a more Westernized lifestyle (mainly from North America, Europe, and Asia) are found in the urban cluster (P=0.001, $R^2 = 0.36$, PERMANOVA test). Ellipses enclose samples within 2 standard deviations for each lifestyle.

C

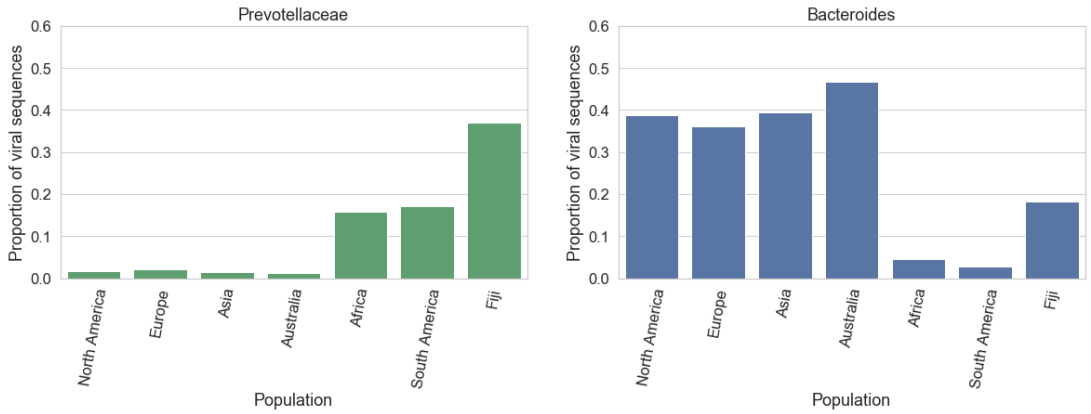


Figure 5.2. Human lifestyle is associated with global gut distribution of phageome types. C) The proportion of VCs that match Prevotellaceae hosts in traditional societies is higher than that of industrialized populations. Conversely, *Bacteroides* hosts are more common in industrialized populations than in traditional societies. Taken together, this result suggests that the composition of the gut phageome at a global scale is driven by the bacterial composition.

5.2.3 Phage carriage across continents

Next, I sought to determine differences in phage carriage according to geographic location (Figure 5.3). It was interesting that despite the large viral diversity that the gut can harbour (21,012 VCs), I detected fewer than 150 VCs in most samples. This threshold could be a result of niche saturation that might prevent exogenous phages from establishing in the gut, mirroring the colonization resistance effect seen in the bacterial gut microbiome. Indeed, longitudinal studies have shown that the gut virome is very stable within individuals (Shkoporov et al., 2019). I did not find significant differences in phage richness across continents except in Africa which had significantly higher diversity.

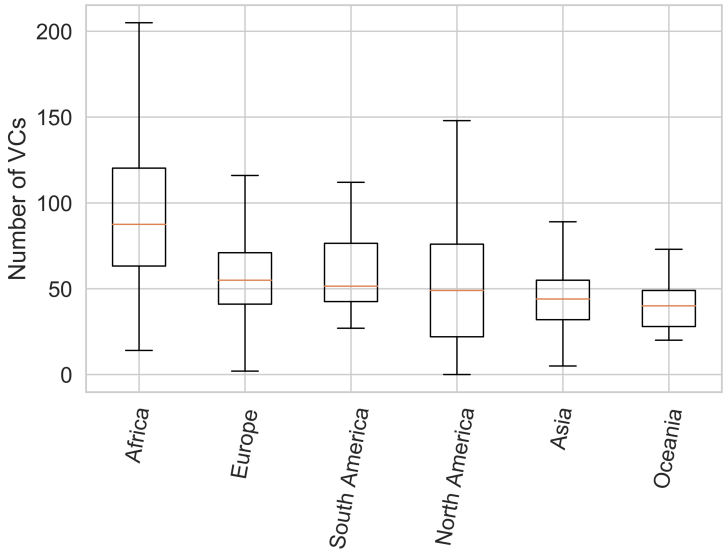


Figure 5.3. Phage carriage across continents. Intra-sample diversity is relatively low compared to the total gut phage diversity. Phage carriage is similar on average per sample across continents except for Africa which is significantly higher. Africa vs Europe (P = 1.82 x 10^{-12}, Mann-Whitney U test), Africa vs South America (P = 0.00033, Mann-Whitney U test), Africa vs North America (P = 4.04 x 10^{-14}, Mann-Whitney U test), Africa vs Asia (P = 6.06 x 10^{-22}, Mann-Whitney U test), Africa vs Oceania (P = 1.64 x 10^{-10}, Mann-Whitney U test)

5.2.4 Uncovering most prevalent phage in global human populations

Stratifying by continent provided me with an unprecedent opportunity to uncover the most prevalent phages around the world. In the case of North America, Europe, and Asia, the host range of the top VCs was dominated by the genera *Bacteroides*, *Bacteroides*_B, and *Parabacteroides*. Notably the p-crAssphage (VC_1) was part of the top VCs for all these continents. Since the gut microbiota of Western societies is dominated by *Bacteroides*, it makes sense that the bacterial hosts of many prevalent VCs are genetically related to this genus. In the case of Africa, South America, and Oceania, for the majority of VCs the bacterial host could not be predicted with the exception of *Faecalibacterium* and *Prevotella*. The absence of host prediction for these continents, may be a consequence of uncultured gut bacteria from these understudied regions, thus hindering efforts to use CRISPR spacers matching or prophage assembly linkage. In general, prevalence of individual VCs was ~25%, the higher prevalence found in South America (~41%) and Oceania (32%) could be result of the limited number of samples to calculate them (<35). Phage prevalence is also dependent on the taxonomic level at

which it's being studied. VCs correspond to subgenus level, however when phages are grouped at genus or family levels their prevalence could substantially increase.

A general observation is that for all continents, phage prevalence follows a power law (Figure 5.4). That is, it appears that across all human populations, there are a few phage clades that are widespread, and they are followed by other clades with decreasing prevalence. Since the rate at which prevalence decreases is proportional to the rank, this behaviour gives rise to a long tail of rare phage clades. High phage prevalence such as that of crAssphage, can be explained by a high prevalence of its bacterial host, while rare phages could be result of them preying on uncommon gut bacteria.

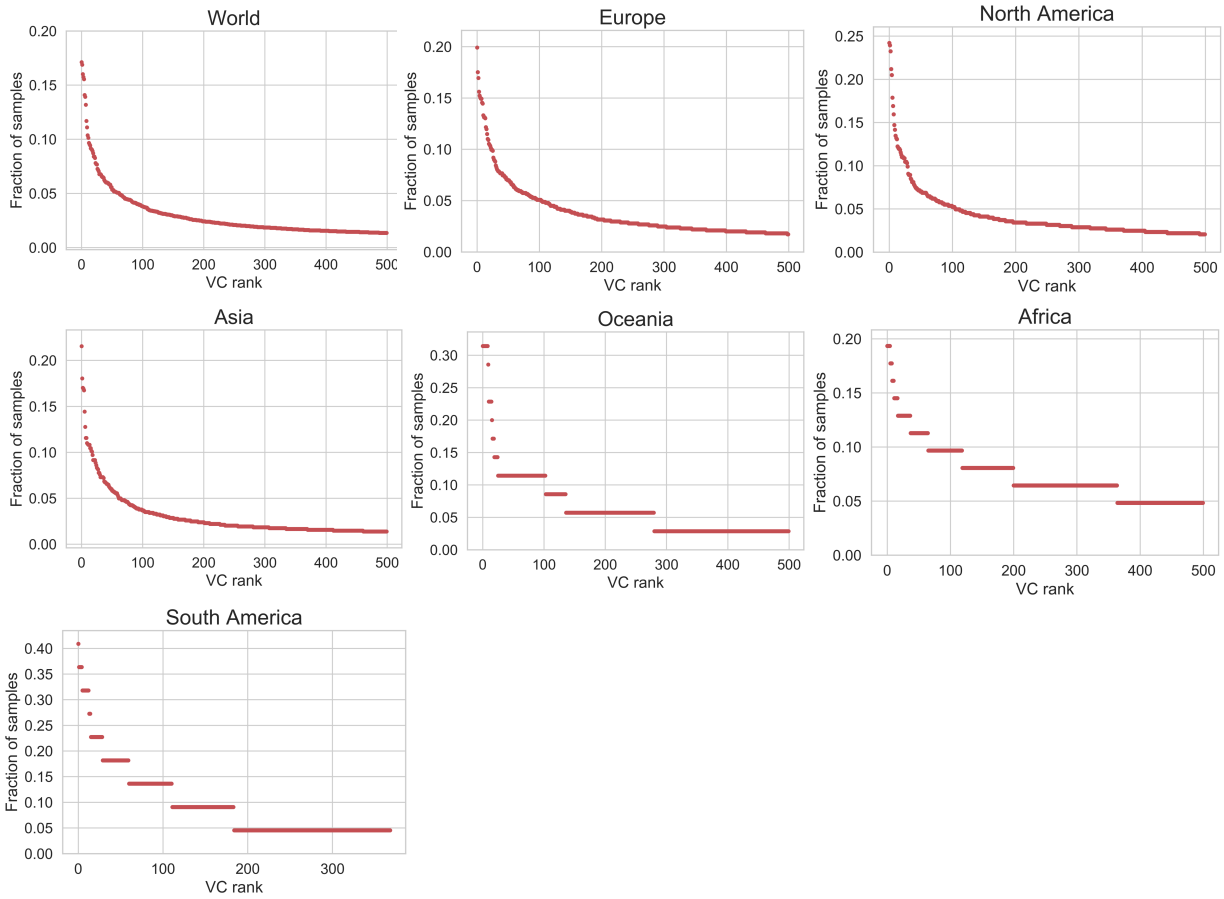


Figure 5.4. Rank prevalence curve for VCs. Prevalence for individual VCs follows a power law distribution across all continents. Phages are usually not found infecting more than ~25% of samples from a given region.

5.2.5 Global distribution of 280 dominant human gut phages

If the gut phageome is predominantly shaped by the bacterial composition, we would expect to observe strong correlation between the prevalence of VCs with that of their bacterial hosts. A clear example is the crAss-like family of gut phages which can be divided into 10 phage genera (Guerin et al., 2018). Genus I, which has been found in a large fraction of Western microbiome samples is able to infect species from the *Bacteroides* genus. In contrast, genera VI, VIII and IX were previously found to be the most prevalent crAss-like phage among Malawian samples (Guerin et al., 2018). Here, I predict that the most probable host of these three phage genera is *Prevotella copri* (rest of crAss-like family predicted hosts in Table 1). In accordance with the results from the Malawian samples, I also found the prevalence of genera VI, VIII and IX to be higher than genus I in Africa and South America (Figure 5.5A). Thus, the crAss-like family is globally distributed with distinct global distribution patterns at the genera level, which appears to be strongly influenced by human lifestyles and enterotypes.

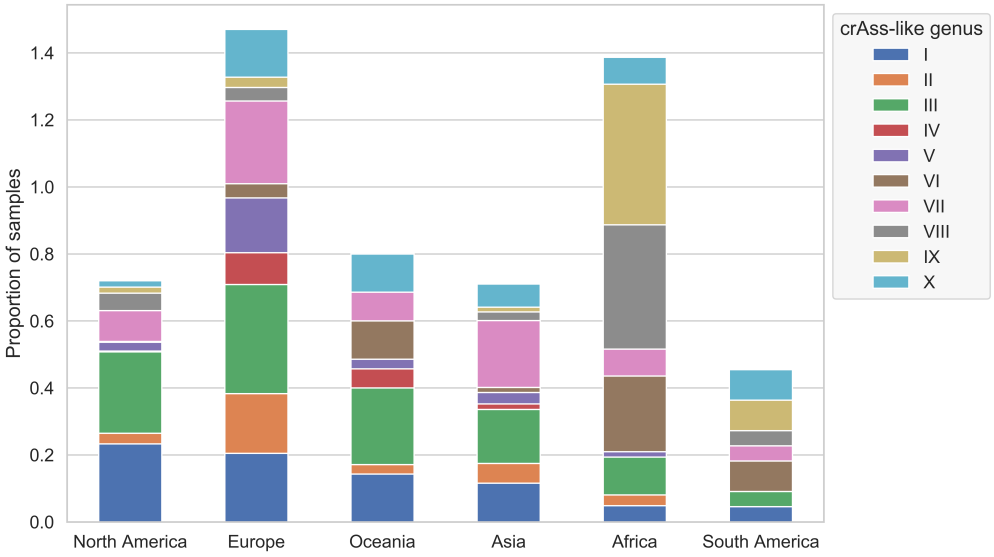
I further investigated if I could identify other gut phage VCs with global distributions. By extending the analysis to all the VCs I was able to detect a total of 280 VCs that were globally distributed (found in at least 5 continents). This represents ~1.3% of all defined VCs (280/21,012). For 119 out of the 280 VCs (42.5%), I was able to classify them to the *Caudovirales* order, whereas the remaining 57.5% remained unclassified. Thus, the majority of globally distributed VCs are completely novel. When I looked at viral families detected within the *Caudovirales*, I detected *Podoviridae* (10 VCs), *Myoviridae* (28 VCs), *Siphoviridae* (43 VCs), and the newly formed family *Herelleviridae* (1 VC). In addition, when I examined at the phage subfamily level, the most common hits corresponded to the *Picovirinae* and *Peduovirinae* subfamilies with 4 VCs each. Importantly, the genomes of 131 members of 57 globally distributed VCs were mined directly from genomes of cultured isolates, providing unique opportunities for follow-up experiments in the lab.

A bacteria-phage network of globally distributed VCs (Figure 5.5B) revealed that *Prevotella* was the most targeted genus (37 VCs), followed by *Faecalibacterium* and *Roseburia* with 15 VCs each. In addition, I observed that in contrast to the Bacteroidales and Oscillospirales, the global VCs associated to the Lachnospirales were highly shared between different genera (Figure 5.5C). Notably, whilst 12 globally distributed VCs were members of the crAss-like family (in black), I was only able to assign a host to 6 VCs which targeted Bacteroidales

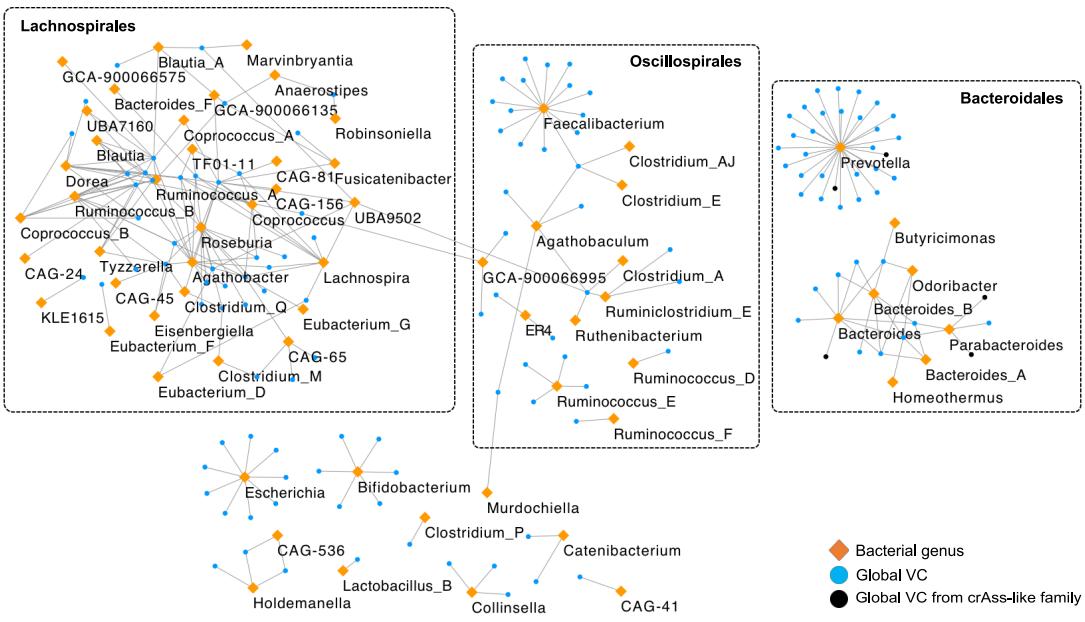
bacteria. I observed that globally distributed phages had a significant broader range (across different genera) than phages found in single continents (P = 1.62x10^{-5}) (Figure 5.5D). This result suggests that broad host-range of certain VCs likely contribute to their expansion across human populations.

Thus, I show that along with 12 crAss-like VCs, there exists a set of at least 280 VCs which are globally distributed. Functional characterization of members of this set will prove useful to shed light on what makes a gut phage to become widespread across human populations.

A



B



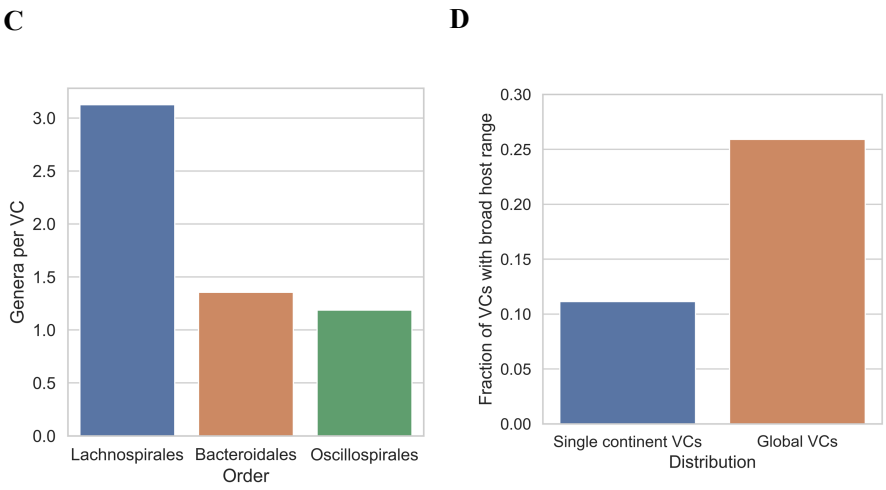


Figure 5.5. Global gut phage clades and their bacterial hosts. A) The crAss-like family is a globally distributed phage. Genera VI, VIII and IX which are predicted to infect a *Prevotella* host are more common in Africa and South America in contrast to genus I which infects a *Bacteroides* host. **B)** Host-phage network of globally distributed VCs (orange) reveals that *Prevotella*, *Faecalibacterium*, and *Roseburia* are the most targeted bacterial genera. VCs that belong to the crAss-like family are highlighted in black; These were predicted to infect *Prevotella*, *Bacteroides*, and *Parabacteroides*. **C)** In contrast to the Bacteroidales and Oscillospirales, the VCs from the Lachnospirales are highly shared. Lachnospirales vs Bacteroidales (P = 9.99 x 10^{-6}, χ^2 test). Lachnospirales vs Oscillospirales (P = 6.55 x 10^{-6}, χ^2 test). **D)** Globally distributed phages had a significantly broader range (above genus) than phages found in single continents (P = 1.63 x 10^{-5}, χ^2 test).

5.2.6 Investigating the concept of a core-virome

Marinque et al. proposed that despite the high interpersonal variation found in the human gut phageome there exists a set of shared phages across individuals (>50%) referred to as the core phageome (Manrique et al., 2016). It was hypothesized that the core phageome is composed of a set of phages which play an important role in maintaining gut microbiome structure/function and thus contribute significantly to human health.

As I showed in Figure 5.4, none of the VCs reached a prevalence >50%, precluding the idea of a core phageome in this work. Nonetheless, I wondered if I could find a reduced set of VCs

that could cover the majority of samples (Figure 5.6A). That is, a sample would be considered covered if at least 1 VC from this set was detected in it. What I found is that at the worldwide level at least one out of 150 VCs were already found in more than 90% of all the samples, and at only 50 VCs the fraction of covered samples was >80% causing the curve to start to plateau. Stratification by continent revealed similar saturation kinetics. At least one out of 50 VCs were found in >50% of samples with the exception of South America (~40%). The more flattened curve observed in South America could be due to the smaller phage genetic diversity captured by GPD. An explanation of why this reduced set of VCs exists is that common phages in the human gut should prey on prevalent bacteria. Certainly, host range prediction of the top 50 VCs for which at least 1 VC is found in >50% of worldwide samples, reveals that these phages infect mostly genera from *Bacteroides*, *Roseburia*, *Parabacteroides*, *Bacteroides_B*, and *Coprococcus*.

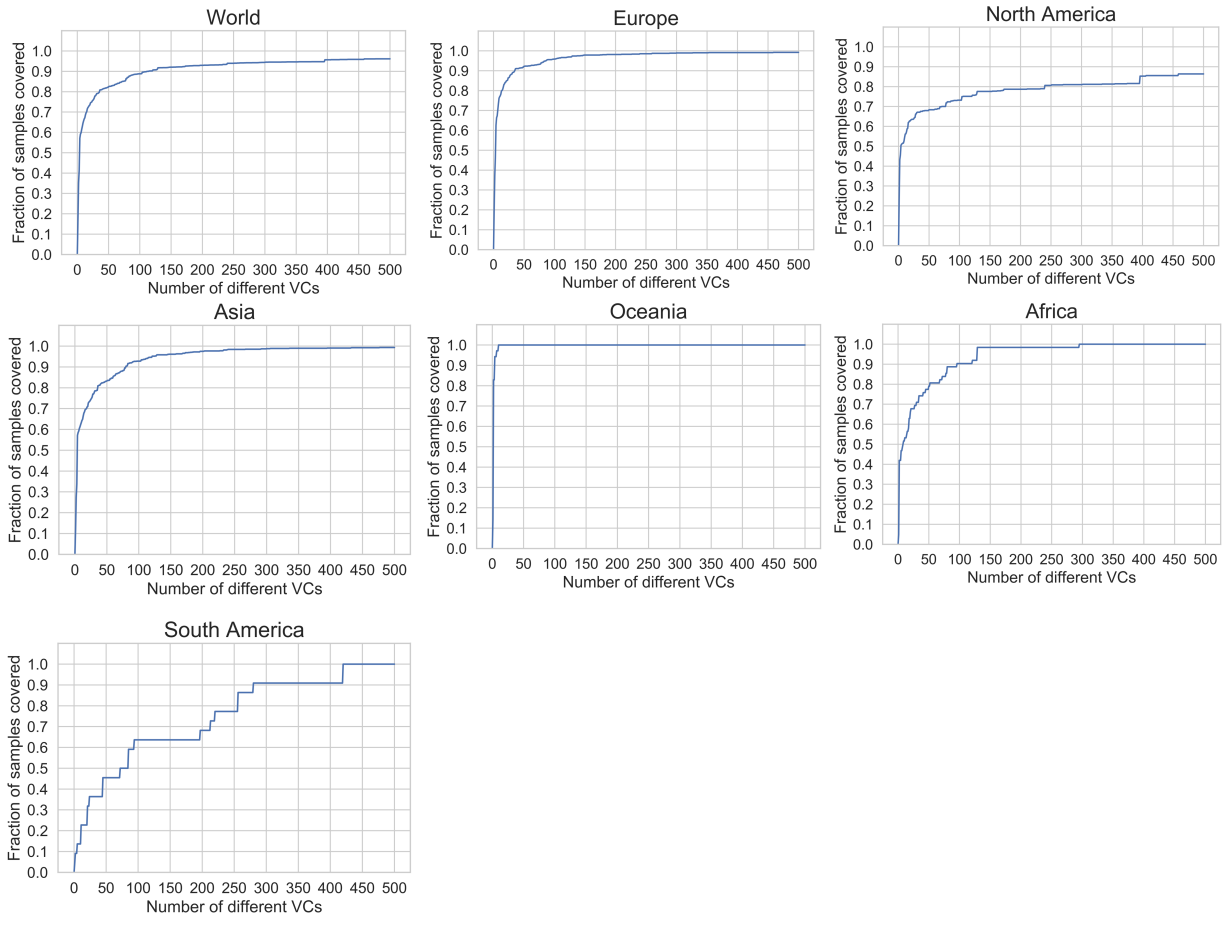
It's also important to mention that although a core virome is unlikely to exist at the ~genus viral level, this finding doesn't reject the idea of highly prevalent viral clades at higher taxonomic ranks. I investigated this idea by measuring the prevalence of the crAss-like family, Gubaphage clade, and *Picovirinae* subfamily across different continents. As we can see in Figure 5.6B, when I pool all the 10 different crAss-like genera, prevalence surpasses ~30% across all continents except in South America, and notably Europe and Africa reach ~70% prevalence. On the other hand, the Gubaphage clade is found well below 20% prevalence across continents, and absent in South America. Europe is the exception with ~40% of samples harbouring a Gubaphage. Finally, I detected the *Picovirinae* subfamily in at least 50% of all samples. Thus, the *Picovirinae* subfamily can be considered a core human phage clade. Notably, its prevalence reaches ~80% in Europe, Africa, and South America. The high prevalence of *Picovirinae* in the last two continents is particularly interesting given that the gut microbiome from African and South American individuals is largely understudied, and thus this finding represents a step forward in understanding and identifying important phages that inhabit their gut.

Analogous to the previous analysis in which I calculated the cumulative fraction of samples covered by each new additional VC, Figure 5.6C shows the same exercise with the crAss-like family, the Gubaphage clade, and the *Picovirinae* subfamily. Combination of the crAss-like family with the Gubaphage clade essentially leaves unchanged the fraction of samples covered when only the crAss-like family is considered, indicating a high co-ocurrence. On the other

hand, when the crAss-like family is combined with the *Picovirinae* subfamily, prevalence surpasses 60% for all continents except in North America (~55%). Notably, Europe and South America reach ~85% prevalence, while in Africa 90% of samples are covered. Combination of the 3 phage clades, does not change much the fraction of samples covered due to the crAss-like and Gubaphage correlation.

Despite only finding one instance of a human core phage (*Picovirinae*), or two if we consider >30% prevalence (*Picovirinae* and crAss-like family), I believe that a proper core phageome may exist. The reason why many studies fail to detect it is because they dereplicate at 95% nucleotide identity. This dereplication threshold is too stringent and thus gives rise to an extremely large variability of the gut phageome (Figure 5.6D). If dereplication was carried out at the level of shared protein clusters (PCs) (e.g. >20% shared PCs), then phage genomes could be clustered at higher phylogenetic levels (genus or subfamily) and phage variation could start to stabilize. Conversely, clustering genomes at very high phylogenetic levels (e.g. order) could result in an unspecific signal.

A



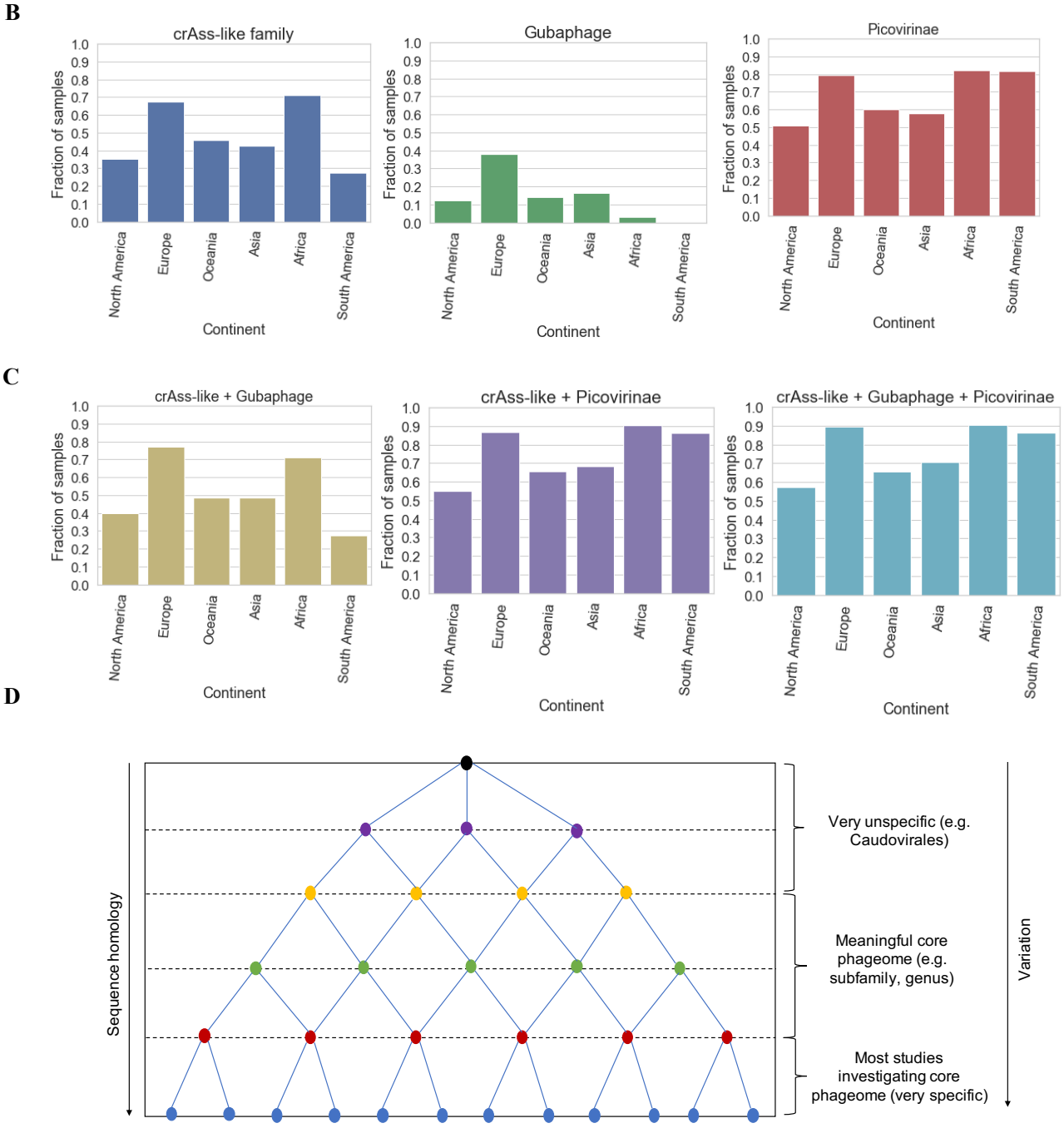


Figure 5.6. Investigating the concept of a core-virome. A) A limited number of VCs are found at least once in a large fraction of human samples across continents. **B)** Analysis of prevalence at higher taxonomic phage clades. CrAss-like phages are found in >30% of worldwide samples, whereas the *Picovirinae* subfamily is found in >50% of samples. **C)** Prevalence analysis with different combinations of the crAss-like family, *Picovirinae*, and Gubaphage clade. **D)** A core phageome may exist, however studies use very stringent dereplication (e.g. 95% nucleotide identity). Probing for higher taxonomic groups may reveal more conserved phages across individuals.

5.3 Conclusions

In this chapter, I analysed the worldwide prevalence and epidemiology of human gut phages by read mapping GPD predictions to a global dataset of human gut metagenomes. This dataset consisted of 3011 samples and spanned all six major continents (Africa, Asia, Europe, North America, South America and Oceania) and 23 countries. The original number of metagenomes considered for this analysis was much bigger (28,060), however samples with a sequencing depth below 50 million reads/sample were removed, as below this threshold I observed a positive correlation between sample depth and number of viral genomes detected. This should be an important consideration for future metagenomic studies of the gut phageome.

I began by studying global patterns of the human gut phageome. A key finding was that urbanization is associated with the composition of the gut phageome. Specifically, when I visualized the distribution of samples, North American, European, and Asian samples segregated from African and South American samples. Samples from the last two continents were derived from communities with non-Western lifestyles. Country-wise stratification showed that Australia belonged to the Western cluster, while Fiji to the rural one. Notably, samples from both countries shared the same lifestyle of their respective cluster. These observations supported the hypothesis that lifestyle, particularly urbanization, may drive differences in the gut phageome across different human populations. In addition, host range prediction of the VCs mapped to each sample, aligned with the expected bacterial enterotype from each continent. Given the correlation of bacterial enterotypes and phageome types, these findings provide evidence that human lifestyle drives global patterns of gut phageomes by mediating changes in the bacterial gut microbiome. Finally, I compared the number of detected VCs per sample across continents. Despite the unprecedented phage diversity found in all samples, I discovered that in general, the majority of individuals only harboured less than 150 VCs.

I then focused on the distribution of individual VCs. A key question was whether there was a set of highly prevalent phage clades which were found across all human populations. For instance, when the p-crAssphage was reported to be found in the majority of analysed samples, a natural question was whether p-crAssphage was a universal highly prevalent phage or if it was exclusive of Western samples. I found that depending on the continent, the most prevalent

phages differed. I found that in North America, Europe, and Asia, p-crAssphage was highly prevalent, but that was not the case for Africa and South America. Nonetheless, for the latter two, I did detect highly prevalent phages that were members of the crAss-like family with a *Prevotella* host range. Despite the dependency of phages on the bacterial composition, I screened for VCs that could be found in all continents. I discovered 280 VCs that were detected in at least 5 continents; a host-phage network showed that the top bacterial genera targeted by these globally distributed VCs were *Prevotella*, *Faecalibacterium*, and *Roseburia*.

The concept of a core virome has sparked controversy in the field, thus I assessed how well it fitted with my data. On one hand, prevalence of individual VCs never reached more than ~25% precluding the idea of a core set of phages shared by at least 50% of individuals. On the other hand, I found that at a worldwide level, at least one of 150 VCs was already found in ~90% of the samples. At the level of continents, at least one of 50 VCs were found in ~50% of the samples. This set of phages is technically not a core virome, but it's surprising the large fraction of samples a relatively small set of VCs can cover given the high level of inter-personal variation found in the gut phageome. A reason why a core virome has not be found may be because analyses are carried out at a very low taxonomic level (e.g. viral species). When I analysed the prevalence of phage clades at a higher taxonomic level, I detected that at least 30% of samples were carrying a crAss-like phage, whereas the *Picovirinae* subfamily was detected in at least 50% of all samples.