Analysis of Different Size Fractions Provides a More Complete Perspective of Viral Diversity in a Freshwater Embayment

Christine N. Palermo,^a Dylan W. Shea,^b Steven M. Short^{a,b,c}

^aDepartment of Physical and Environmental Sciences, University of Toronto Mississauga, Mississauga, Ontario, Canada

^bDepartment of Ecology and Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada

^cDepartment of Biology, University of Toronto Mississauga, Mississauga, Ontario, Canada

ABSTRACT Inspired by recent discoveries of the prevalence of large viruses in the environment, we reassessed the longstanding approach of filtering water through small-pore-size filters to separate viruses from cells before metagenomic analysis. We collected samples from three sites in Hamilton Harbour, an embayment of Lake Ontario, and studied 6 data sets derived from $<0.45\text{-}\mu\text{m}$ - and $>0.45\text{-}\mu\text{m}$ -size fractions to compare the diversity of viruses in these fractions. At the level of virus order/family, we observed highly diverse and distinct virus communities in the $>0.45\text{-}\mu\text{m}$ -size fractions, whereas the $<0.45\text{-}\mu\text{m}$ -size fractions were composed primarily of *Caudovirales*. The relative abundances of *Caudovirales* for which hosts could be inferred varied widely between size fractions, with higher relative abundances of cyanophages in the $>0.45\text{-}\mu\text{m}$ -size fractions, potentially indicating replication within cells during ongoing infections. Many viruses of eukaryotes, such as *Mimiviridae*, *Phycodnaviridae*, *Iridoviridae*, and *Poxviridae*, were detected exclusively in the often-disregarded $>0.45\text{-}\mu\text{m}$ -size fractions. In addition to observing unique virus communities associated with each size fraction from every site we examined, we detected viruses common to both fractions, suggesting that these are candidates for further exploration because they could be the product of ongoing or recent lytic events. Most importantly, our observations indicate that analysis of either fraction alone provides only a partial perspective of double-stranded DNA (dsDNA) viruses in the environment, highlighting the need for more comprehensive approaches for analyzing virus communities inferred from metagenomic sequencing.

IMPORTANCE Most studies of aquatic virus communities analyze DNA sequences derived from the smaller-size “free-virus” fraction. Our study demonstrates that analysis of virus communities using only the smaller-size fraction can lead to erroneously low diversity estimates for many of the larger viruses such as *Mimiviridae*, *Phycodnaviridae*, *Iridoviridae*, and *Poxviridae*, whereas analyzing only the larger- $>0.45\text{-}\mu\text{m}$ -size fraction can lead to underestimates of *Caudovirales* diversity and relative abundance. Similarly, our data show that examining only the smaller-size fraction can lead to underestimations of viroplasm and cyanophage relative abundances that could, in turn, cause researchers to assume their limited ecological importance. Given the considerable differences we observed in this study, we recommend cautious interpretations of environmental virus community assemblages and dynamics when based on metagenomic data derived from different size fractions.

KEYWORDS bacteriophages, dsDNA viruses, eukaryotic viruses, metagenomics, microbial communities, viral ecology, viroplankton

Environmental viruses play key roles in shaping microbial communities and structuring ecosystems vital to the health of the planet and our species. They act as major drivers of microbial community succession (1, 2), influence global biogeochemical

Citation Palermo CN, Shea DW, Short SM. 2021. Analysis of different size fractions provides a more complete perspective of viral diversity in a freshwater embayment. *Appl Environ Microbiol* 87:e00197-21. <https://doi.org/10.1128/AEM.00197-21>.

Editor Christopher A. Elkins, Centers for Disease Control and Prevention

Copyright © 2021 American Society for Microbiology. All Rights Reserved.

Address correspondence to Christine N. Palermo, christine.palermo@mail.utoronto.ca, or Steven M. Short, steven.short@utoronto.ca.

Received 27 January 2021

Accepted 11 March 2021

Accepted manuscript posted online 19 March 2021

Published 11 May 2021

cycles (3), and may play important roles in the evolution of eukaryotes (4, 5). Metagenomics has helped uncover a remarkable and diverse array of previously undiscovered viruses, leading to the rapid expansion of virus sequence databases (6–14). As new analytical tools and pipelines are developed, constant exploration and scrutiny of new and existing bioinformatic and methodological approaches are necessary when addressing primary questions in environmental virology.

Early metagenomic studies of aquatic virus communities were conducted using DNA extracted from water samples passed through filters with 0.2- μm or smaller pore sizes (15, 16). More recently, the discoveries of viruses with increasingly larger capsid sizes (as reviewed in reference 17) necessitated reevaluation of common filtration methods when attempting to capture entire aquatic virus communities. It is now clear that the $<0.2\text{-}\mu\text{m}$ -size fraction excludes many diverse larger viruses, prompting some researchers to use 0.45- μm -pore-size filtration in attempt to sample these larger viruses and obtain a more complete picture of the total virus community (18–21). For example, in one study of soils, use of 0.45- μm -pore-size filters approximately doubled the number of observable virus-like particles compared to that with the 0.2- μm -pore-size filters (22). Thus, the choice of filtration schemes to enrich aquatic samples for viral community analysis is nontrivial and is exacerbated by the presence of larger viruses, particularly those in the *Mimiviridae* family, that can be captured on 0.45- μm -pore-size filters. Many of these larger *Mimiviridae* can be excluded from analyses of total virus communities when examining filtrates from 0.22- μm or 0.45- μm filtration, with the potential for slightly higher mimivirus recovery with 0.45- μm -pore-size filters (22, 23). Additionally, viruses within or attached to larger bacterial cells, eukaryotic cells, or particulate material are likely to be captured on the filters and excluded from the analysis of filtrates used to obtain a picture of the viral community. Some cell-associated viruses may pass through 0.22- μm or 0.45- μm filters and could be lost during the viral DNA extraction process if they are within ultra-small bacteria that are common in aquatic environments (24).

We previously studied virus communities in Hamilton Harbour (14), an urban eutrophic freshwater embayment of Lake Ontario, Canada, that is well studied due to its economic importance and long-term remediation efforts. We mined viral sequences from metagenome libraries derived from $>0.22\text{-}\mu\text{m}$ -size fractions originally collected to study cellular communities. Typically, to avoid confounding viral and cellular sequences, viral sequences within this size fraction are filtered out and, therefore, are not included in viral community analyses (25–28). Our study revealed diverse, seasonally dynamic viral communities that were distinct at different sites within the harbor. In most samples, virophages and *Mimiviridae* were highly abundant, contrasting other studies of similar environments that focused on viral communities in the $<0.22\text{-}\mu\text{m}$ -size fraction. This prompted us to examine both $<0.45\text{-}\mu\text{m}$ - and $>0.45\text{-}\mu\text{m}$ -size fractions collected from the same water samples in attempt to capture the entire virus community and to evaluate the diversity and relative abundances of different groups of viruses partitioning into each fraction. To achieve this objective, we used two common methods for targeting aquatic virus communities: (i) water filtration through 0.45- μm -pore-size filters and concentration of viruses in the filtrate using iron chloride flocculation (29), followed by DNA extraction and sequencing of liberated viruses, and (ii) mining viral DNA from cellular metagenomes, i.e., DNA extracted and sequenced from the material captured on 0.45- μm -pore-size filters. The resulting size-fractionated double-stranded DNA (dsDNA) viral communities were then analyzed using established metagenomics techniques to address the hypothesis that the virus communities observed in each fraction would differ, as each fraction would include unique viral taxa.

(Data presented here were included in a manuscript submitted to a preprint archive [30].)

RESULTS

Virus community composition in $<0.45\text{-}\mu\text{m}$ - and $>0.45\text{-}\mu\text{m}$ -size fractions. Contigs annotated as viruses comprised 15% to 20% of all contig annotations in the $<0.45\text{-}\mu\text{m}$ -size

FIG 1 Relative abundances of all virus taxa from the <0.45- μm -size fractions and >0.45- μm -size fractions at each site. The table above the bars shows the number of virus contigs that comprise each bar and the percentage of all contigs that were annotated as viral for each sample. Sample similarity is represented by the Bray-Curtis dissimilarity dendrogram, where blue branches represent <0.45- μm -size fraction samples and red branches represent >0.45- μm -size fraction samples.

fractions, while fewer than 1% of annotated contigs were of viral origin in the >0.45- μm -size fractions (Fig. 1). Diverse virus contigs were observed in both size fractions belonging to the virus groups *Caudovirales*, virophages (*Lavidaviridae*), *Mimiviridae*, *Phycodnaviridae*, other bacterial viruses, other dsDNA viruses, single-stranded DNA (ssDNA) viruses, and unclassified viruses. The viral communities in <0.45- μm -size fractions were similar at the 3 sites and clustered closely together in the Bray-Curtis dissimilarity dendrogram (Fig. 1). While many virus groups were found in both size fractions, most were present at only low relative abundances in the <0.45- μm -size fractions; *Caudovirales*, other bacterial viruses, and unclassified viruses comprised >99% of all virus contigs in the smaller-size fractions. *Caudovirales* relative abundances were remarkably consistent in the smaller-size fractions at 82% but were more variable in the larger-size fractions ranging from 2% to 42%. Additionally, there was more variability in the relative abundances of viral groups between sites in the >0.45- μm -size fractions than the <0.45- μm -size fractions (Fig. 1).

Relative abundances of virophages in >0.45- μm -size-fraction samples ranged from 27% to 79%, whereas the relative abundance of *Mimiviridae* ranged from 10% to 14% (Fig. 1). In contrast, virophages and *Mimiviridae* made up <0.1% at all 3 sites in the <0.45- μm -size fraction (Fig. 1). Algal viruses in the family *Phycodnaviridae* were present in all samples but represented a small fraction of the viral community. In <0.45- μm -size fractions, *Phycodnaviridae* constituted 0.1% of the virus community at all 3 sites and were slightly more abundant in the >0.45- μm samples ranging from 1% to

FIG 2 Relative abundances of *Caudovirales* families from the <0.45-μm-size fractions and >0.45-μm-size fractions at each site. Numerical values above the bars are the percentage of *Caudovirales* in the entire virus community for each sample. Sample similarity is represented by the Bray-Curtis dissimilarity dendrogram, where blue branches represent <0.45-μm-size-fraction samples and red branches represent >0.45-μm-size-fraction samples.

2% relative abundance. Relative abundances of unclassified viruses ranged from 15% to 16% in the <0.45-μm-size fractions and from 6% to 16% in the >0.45-μm-size fractions (Fig. 1).

***Caudovirales* community composition and host inference.** Viruses belonging to the order *Caudovirales* were predominated by members of the *Siphoviridae* family in all samples (Fig. 2). In <0.45-μm samples, the *Siphoviridae* comprised >88% of all *Caudovirales* at the 3 sites, while the relative abundance of *Siphoviridae* ranged from 54% to 76% in >0.45-μm-size-fraction samples (Fig. 2). *Myoviridae* was the second most abundant family within *Caudovirales* in the >0.45-μm-size fractions, making up 14% to 36% of their respective viral communities. In the <0.45-μm-size fractions, *Podoviridae* was the second most abundant *Caudovirales* family making up 4% to 5% of *Caudovirales* in their respective communities (Fig. 2). Overall, *Podoviridae* were present at low relative abundances in every sample, but with slightly higher values in the <0.45-μm-size fractions. As shown in the Bray-Curtis dissimilarity dendrogram, community composition was similar among sites in the <0.45-μm-size fraction, whereas in the larger-size fraction, only sites 1 and 2 clustered together; site 3 was more similar to the smaller-size fractions than to the other larger-size fractions (Fig. 2).

Phages of *Methylophilaceae*, *Methylophilales*, and *Pontimonas* were most abundant in the <0.45-μm-size fractions, together comprising 53% to 60% of all *Caudovirales* with inferred hosts (Fig. 3). L5 viruses that infect *Mycobacterium* and *Rhodococcus* species were the most abundant phages of heterotrophic bacteria in the >0.45-μm-size fractions at sites 1 and 2. Phages inferred to infect *Lactococcus* were the most

FIG 3 Bubble plot showing relative abundances of *Caudovirales* for which hosts could be inferred with abundances denoted by the size of the bubble. Shading represents the percentage of all *Caudovirales* that each bubble represents. Sample similarity is represented by the Bray-Curtis dissimilarity dendrogram, where blue branches represent <0.45- μm -size-fraction samples and red branches represent >0.45- μm -size-fraction samples.

abundant phages in the >0.45- μm -size fraction at site 3, comprising 16% of all *Caudovirales* annotations in that sample. Of the inferred cyanophages, phages of *Synechococcus* were the most abundant in the <0.45- μm -size fractions and in the >0.45- μm -size fraction at site 3, whereas the >0.45- μm -size fractions at sites 1 and 2 were dominated by phages of *Microcystis*. Phages of *Aphanizomenon* were detected in every sample, with higher relative abundances in the larger-size fractions than in the smaller-size fractions. Across all samples, the “other bacteriophages” category contained 85 unique phages, many of which were viruses of heterotrophic bacteria. In the Bray-Curtis dissimilarity analysis, distinct clusters formed for the <0.45- μm -size fractions

FIG 4 Unique and shared elements in the samples pooled into <0.45- μm - and >0.45- μm -size fractions. (A) Venn diagram of the number of unique contig annotations in each size fraction. Bar charts show the relative abundance of virus groups in each category of the Venn diagram. (B) Venn diagram of the number of unique inferred *Caudovirales* hosts in each size fraction.

and the >0.45- μm -size fractions. For both size fractions, *Caudovirales* communities from sites 1 and 2 clustered together, and the communities at site 3 were less similar to those at sites 1 and 2.

Virus diversity in the <0.45- μm - and >0.45- μm -size fractions. Of 694 unique contig assignments for sequences generated from all 6 samples, 93% were detected in at least one sample from the <0.45- μm -size fractions, whereas only 24% were detected in at least one of the >0.45- μm -size fractions (Fig. 4A). Seventy-six percent of unique contig assignments were found only in the <0.45- μm -size fractions, while 7% of unique contig assignments were found only in the >0.45- μm -size fractions. The two size fractions shared 17% of unique contig assignments, most of which were *Caudovirales*. Ninety-one percent of the unique contig assignments that were identified solely in the <0.45- μm -size fractions were *Caudovirales* compared to 35% in the >0.45- μm -size fractions. The majority of unique *Caudovirales* and other bacterial viruses were detected only in the <0.45- μm -size fractions, whereas *Iridoviridae* and *Poxviridae* were only detected in the >0.45- μm -size fractions. In addition to the 11 unique *Mimiviridae* contigs detected in both size fractions, 7 unique *Mimiviridae* contigs were detected in the <0.45- μm -size fractions, and 11 unique *Mimiviridae* contigs were detected in the >0.45- μm -size fractions. There were 18 unique *Phycodnaviridae* contigs detected only in the <0.45- μm -size fraction compared to 6 unique *Phycodnaviridae* contigs detected only in the >0.45- μm -size fraction. Seven unique

virophage contigs were detected in both size fractions, and an additional 3 were detected in the <0.45- μ m-size fractions.

There were 86, 87, and 89 unique *Caudovirales* contigs with inferred hosts in the smaller-size fractions at sites 1, 2, and 3, respectively. We observed lower richness in the larger-size fractions, which contained 15, 31, and 12 unique *Caudovirales* with inferred hosts at sites 1, 2, and 3, respectively. All inferred *Caudovirales* hosts except *Salinibacter* were captured in the <0.45- μ m-size fractions (Fig. 4B). All *Caudovirales* with inferred hosts with >1% relative abundance (i.e., those in Fig. 3) were captured by both size fractions, except *Pontimonas* and *Pseudoalteromonas* phages, which were not detected in the larger-size fractions. Many of the 85 viruses with inferred hosts within the “other bacteriophages” category were only detected in the smaller-size fraction and at low relative abundances.

DISCUSSION

Data considerations and assumptions. Beyond factors related to data processing programs and pipelines, there are considerations and assumptions related to our methods and results that merit some discussion. The sequencing methods used in this study only capture DNA viruses; thus, the dsRNA and ssRNA virus communities were not considered. Furthermore, sequencing would only capture ssDNA viruses in dsDNA form, for instance, during replication inside their hosts. This is reflected in our results, which showed that most ssDNA viruses were detected in the larger->0.45- μ m-size fractions that would include viral DNA within cells. Therefore, the data and conclusions presented in this study apply almost exclusively to the dsDNA virus communities.

Compared to prokaryote and eukaryote sequences, virus sequences are underrepresented in reference databases, influencing the reported diversity of viruses in metagenomes (discussed in detail in reference 14). This inevitably leads to a large portion of sequences in environmental metagenomes remaining unclassified and being discarded at the annotation step in the data processing pipeline. To increase the likelihood of annotating virus sequences, we used the NCBI-nonredundant (nr) database which encompasses reference sequences from both curated and noncurated data sources. Of the virus sequences available in reference databases, some groups are more comprehensively represented than others. For example, *Caudovirales* genomes are 2 orders of magnitude more numerous than virophage genomes in NCBI RefSeq (data from <https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi>). This leads to a higher likelihood of annotating *Caudovirales* sequences as originating from discrete taxa than virophage sequences. Our observations of disproportionately high virophage abundances compared to those in reference databases aligns with previous observations of their high relative abundances in larger-size fractions (14, 31) and reinforces the notion that sequence annotations are only as good as the databases from which they are derived.

In this study, we inferred the presence and abundance of viruses using the normalized relative abundances of assembled contigs. There are many viral genes present in prokaryotic and eukaryotic genomes due to the intimate evolutionary relationship between viruses and their hosts. For example, it is well known that many bacterial genomes are littered with cryptic prophage sequences (32), and endogenized retrovirus sequences are a common and surprisingly abundant feature of mammalian genomes (33). Unless viral homologues within host genomes are specifically annotated as such, they could be incorrectly annotated as viruses (34). Furthermore, metagenomic data cannot distinguish between viruses that are involved in active infections and inactive viruses that are adsorbed to cells or particles, which could also lead to an overestimation of viral influence. Therefore, a caveat of this study is the fact that we could not distinguish viral DNA integrated into host genomes versus viral DNA packaged within virions.

Comparisons of the <0.45- μ m- and >0.45- μ m-size fractions. We observed major differences between the <0.45- μ m-size fractions and >0.45- μ m-size fractions at almost all levels of taxonomic analysis. DNA extractions from the smaller-size fractions

yielded more DNA than the larger-size fractions. DNA was extracted from only one-quarter of the filters for the larger-size fractions, and DNA yield may have been impacted by incomplete lysis, adsorption to particles, or incomplete recovery of supernatant after bead beating. Sequencing library concentrations were higher in the smaller-size fractions, but despite this, there were no apparent trends in the number of reads generated before or after quality control or in the total number of contigs assigned. The number of virus contigs assigned was much higher in the $<0.45\text{-}\mu\text{m}$ -size fractions than in the $>0.45\text{-}\mu\text{m}$ -size fractions, as expected due to the enrichment of virus particles in the $<0.45\text{-}\mu\text{m}$ -size fractions.

For the sake of brevity, we will use only the virus name or taxonomic group when referring to contigs annotated as those viruses. Across all levels of taxonomic analysis, the smaller-size fractions clustered closely together in the Bray-Curtis dissimilarity dendrogram, indicating few differences in the relative abundances of virus groups between sites. The larger-size fractions had higher Bray-Curtis dissimilarity values at all levels of analysis. At the order and family levels, the smaller-size fractions consisted almost entirely of *Caudovirales*, other bacterial viruses, and unclassified viruses, whereas the larger-size fractions were more diverse, with lower relative abundances of *Caudovirales* but higher relative abundances of virophages, *Mimiviridae*, *Phycodnaviridae*, *Iridoviridae*, *Poxviridae*, other dsDNA viruses, and ssDNA viruses. Virus community composition in the smaller-size fractions is typical of freshwater viromes, which are predominated by *Caudovirales* and other phages. Viruses associated with eukaryotes such as *Mimiviridae* and virophages are often undetected or reported at very low abundances in these viromes (35–41). However, Roux et al. (31) demonstrated that diverse virophage and nucleocytoplasmic large DNA virus (NCLDV) communities can be identified by mining sequences derived from the larger-size fraction that is typically used to study cellular organisms.

At the level of *Caudovirales* families, the larger-size fraction from site 3 clustered more closely with the smaller-size fractions than with the other larger-size fractions. While *Siphoviridae* were the most abundant *Caudovirales* family in every sample, relative abundances of *Myoviridae* in the larger-size fractions were higher and more variable between the sites. Previous metagenomic studies of Lake Ontario and nearby Lake Erie found that *Myoviridae* was the dominant *Caudovirales* family (39). Likewise, metagenomic data from Lake Baikal showed a predominance of *Myoviridae* (35, 42), whereas *Siphoviridae* and *Myoviridae* relative abundances were similar in Lakes Michigan, Pavin, and Bourget (38, 43). Therefore, our findings of highly abundant *Siphoviridae* relative to *Myoviridae* in Hamilton Harbour contrasts with previous studies of *Caudovirales* communities in other freshwater lakes.

Of the *Caudovirales* for which hosts could be inferred, the larger-size fractions from sites 1 and 2 had higher relative abundances of cyanophages than the smaller-size fractions and the site 3 larger-size fraction. These higher relative abundances in the larger-size fractions may indicate active cyanophage infections, as Hamilton Harbour is known to experience algal blooms during the summer and late autumn (44). *Microcystis* phage sequences were particularly abundant in the $>0.45\text{-}\mu\text{m}$ -size fractions from sites 1 and 2, suggesting ongoing active infections of *Microcystis* spp. *Microcystis aeruginosa* is a well-known bloom-forming cyanobacterium in Hamilton Harbour (45), Lake Ontario (46), and diverse freshwater lakes around the globe (47, 48); therefore, high relative abundances of *Microcystis* phages were unsurprising. However, in the $<0.45\text{-}\mu\text{m}$ -size fractions, contribution of *Microcystis* phages to the total cyanophage community was low, demonstrating the different insights into the virus communities that the two size fractions provide. Moreover, *Microcystis* phages were not detected in the $>0.45\text{-}\mu\text{m}$ -size fraction from site 3, reinforcing previous findings of distinct virus communities over relatively small spatial scales in Hamilton Harbour (14). We observed more pronounced differences between the size fractions at site 3, where *Lactococcus* phage sequences were highly abundant in the $>0.45\text{-}\mu\text{m}$ -size fraction but undetectable in the $<0.45\text{-}\mu\text{m}$ -size fraction. Across all samples and *Caudovirales* with known hosts, the *Lactococcus* phages in the larger-size fraction from site 3 were the

most abundant of all *Caudovirales* annotated in that sample. The most abundant *Lactococcus* phage was *Lactococcus* phage 16802, a phage that infects *Lactococcus lactis*, the most abundant lactic acid bacterium in 7 Japanese lakes (49). The discovery of highly abundant *Lactococcus* phages in a eutrophic embayment of the Great Lakes is intriguing given previous reports that *Lactococcus* species may mitigate harmful algal blooms by acting as cyanobacteriicidal bioagents (50). Regardless of their potential ecological interactions with bloom-forming species, the high relative abundance of *Lactococcus* phages is indicative of their ecological importance in this system.

In all <0.45- μm -size-fraction samples, phages infecting *Methylophilales*, *Methylophilaceae*, and *Pontimonas* were the most abundant *Caudovirales* with inferred hosts. *Methylophilaceae* is a ubiquitous and metabolically diverse bacterial family (51), with observed enrichments in the Great Lakes surface waters (52). Members of this bacterial family play important roles in global carbon (53) and nitrogen cycling via methanol oxidation linked to denitrification (54) and seem to be ecologically important in Hamilton Harbour, a nutrient-rich environment that supports high microbial biomass. Therefore, the observation of these phages is interesting as they may indirectly influence nitrogen cycling in Hamilton Harbour by acting as a major source of *Methylophilaceae* mortality. Furthermore, some phages of *Methylophilaceae* carry auxiliary metabolic genes (AMGs) that enhance methanol oxidation by their hosts (55), which may in turn act as a viral control mechanism of denitrification; however, the most abundant *Methylophilaceae* phage in our samples (P19250A) does not appear to contain these AMGs (56). *Methylophilaceae* phage P19250A was originally isolated from the <0.2- μm -size fraction from an oligotrophic freshwater lake but has been detected in the viromes of 8 other diverse freshwater systems, including eutrophic lakes and even Lake Ontario (56). *Methylophilaceae* grow rapidly and have high biomass yields (53), potentially contributing to our observations of high relative abundances of their phages during a major lytic event. Because of their potential relevance to nitrogen cycling in Hamilton Harbour, further study of these apparently abundant phages may be warranted.

The *Pontimonas* genus contains only one described strain, *Pontimonas salivibrio* strain CL-TW6^T, which was recently isolated from a coastal marine environment along with its virus and was hypothesized to occupy environments that experience significant stresses (57), such as Hamilton Harbour. The researchers concluded that its lytic virus was poorly adapted to replicate in this strain, since repeated attempts to sequence its genome were unsuccessful. Given the high abundances of this *Pontimonas* phage in all <0.45- μm -size fractions from Hamilton Harbour, we presume that it has a permissive host supporting its replication, yet its absence in all the >0.45- μm -size fractions warrants further investigation of this virus-host system that may be prevalent in Hamilton Harbour.

The results of this study demonstrated that *Caudovirales* were more effectively captured in the smaller-size-fraction samples, whereas the larger, eukaryote-infecting viruses were predominantly observed in the larger-size fractions. Interestingly, a subset of viruses was detected in both size fractions, suggesting that they are highly abundant and likely ecologically important in this environment. These viruses were sufficiently abundant to be detected as free viruses in the water column as well as in the larger-size fraction, potentially indicating ongoing or recent lytic events. Of particular importance to Hamilton Harbour, we detected a variety of cyanophages in both size fractions, including phages of *Microcystis* (Ma-LMM01 and MaMV-DC), *Prochlorococcus* (P-HM2), *Aphanizomenon* (vB_AphaS-CL131), and *Synechococcus* (ACG-2014, S-PM2, S-CAM7, S-T4, S-CBS2, S-SKS1, S-PRM1, S-CAM22, S-CBS4, S-WAM2, S-CAM8, S-TIM5, and KBS-S-2A). We also detected viruses of eukaryotic primary producers in both size fractions, including *Chrysochromulina ericina* virus, unclassified chloroviruses, Yellowstone Lake phycodnavirus 2, and *Micromonas* sp. RCC1109 virus MpV1. Identifying overlap between the virus communities in different size fractions

provides the impetus for further exploration, because the presence of these viruses in both fractions allows us to hypothesize that they may be the product of recent or ongoing active viral infections in aquatic environments.

Site comparisons and similarities with a previous study of Hamilton Harbour viruses. As noted above, across all levels of analysis and both size fractions, sites 1 and 2 were generally more alike than site 3, and the most noticeable differences between sites were in sequences from the $>0.45\text{-}\mu\text{m}$ -size fractions. These observations reinforce previous observations (14), allowing us to conclude that distinct virus communities can be observed over small spatial scales in Hamilton Harbour. Furthermore, this pattern of community structure appears to be a stable ecological phenomenon rather than a chance singular observation from a snap-shot study of the environment. In 2015, $>0.22\text{-}\mu\text{m}$ -size fractions from Hamilton Harbour were analyzed from samples collected biweekly from July to September from sites 1 and 3. Different methodologies were used to study viruses from larger-size fractions in Hamilton Harbour in 2015 and 2019, including different approaches for water sampling, filtration, and DNA extraction. Furthermore, in these separate studies, samples were collected and processed by different researchers several years apart, and different sequencing machines, raw read lengths, and data analysis parameters were used. Despite these large differences in methodology, in both studies, we observed similar patterns of high viroplage abundances in the larger-size fractions, which were particularly pronounced at site 3. As in the 2015 samples, *Mimiviridae* relative abundances in the $>0.45\text{-}\mu\text{m}$ -size fractions did not fluctuate to the same extent as the virophages, despite their presumably intimate association. Like the study of samples collected in 2015, in the present study of 2019 samples, we observed low relative abundances of *Phycodnaviridae* even though Hamilton Harbour is known to support high algal biomass during the sampling period. Again, this indicates that *Mimiviridae* may be the dominant alga-infecting viruses in this system.

Experimental and ecological relevance. In this study, we opted to use $0.45\text{-}\mu\text{m}$ -pore-size filters rather than $0.2\text{-}\mu\text{m}$ -pore-size filters to maximize the likelihood of capturing large, free virus particles in the filtrate in attempt to distinguish between free viruses in the $<0.45\text{-}\mu\text{m}$ -size fraction and viruses associated with larger particulate material in the $>0.45\text{-}\mu\text{m}$ -size fraction. *Mimiviridae* and other large DNA viruses are an exception to this generalization, as they could be captured on the $>0.45\text{-}\mu\text{m}$ -pore-size filters even if present as free viruses in the water column. Traditional viromics studies that look only at the $<0.2\text{-}\mu\text{m}$ - or $<0.45\text{-}\mu\text{m}$ -size fractions may exclude large portions of the *Mimiviridae*, *Phycodnaviridae*, *Iridoviridae*, and *Poxviridae* communities, potentially leading to an underestimation of their influence on the microbial communities. Additionally, members of the *Caudovirales* community involved in ongoing infections of larger bacterial cells may be overlooked. Conversely, mining viral sequences in the larger-size fractions may not sufficiently capture the diversity of *Caudovirales*, many of which may be present at low abundances but acting as a seed bank (58) for future infection. The $>0.45\text{-}\mu\text{m}$ -size fractions captured the diversity of larger viruses more effectively than the $<0.45\text{-}\mu\text{m}$ -size fractions; however, many viruses in the $>0.45\text{-}\mu\text{m}$ -size fractions may also be missed by sequencing efforts due to the low representation of viruses in these data sets. Larger cellular genomes comprise most of the DNA in the $>0.45\text{-}\mu\text{m}$ -size fractions, leading to low sequencing coverage of viruses and, therefore, likely reducing diversity estimates by only permitting sequencing of the most abundant viruses. While all discrete viroplage taxa were observed in the $<0.45\text{-}\mu\text{m}$ -size fractions, viroplage contigs observed in the $>0.45\text{-}\mu\text{m}$ -size fractions were often longer and more numerous. Therefore, high viroplage relative abundance estimates from the larger-size fractions were not simply a result of low detection of other groups of viruses. The high relative abundance of virophages in the larger-size-fraction samples could be the result of their intimate associations with cellular or giant virus hosts, as demonstrated in Mavirus and Sputnik virophages (59, 60). Whether these examples are widely applicable to viroplage-virus-host systems remains to be assessed, yet the high

relative abundance of virophages in all >0.45- μm -size fractions is indicative of their potential influence on the *Mimiviridae* and their eukaryotic host communities.

These unequivocal results from all samples highlight the potential biases associated with prefiltering water samples to concentrate viruses prior to metagenomic sequencing, demonstrating that the entire virus community cannot be captured in either size fraction alone. While we observed highly diverse and different virus communities in the >0.45- μm -size fractions, the <0.45- μm -size fractions were similar and composed primarily of bacteriophages belonging to the *Caudovirales* family. However, higher relative abundances of cyanophages were observed in the larger-size fractions, potentially indicating replication within cells during ongoing infections. As noted previously, larger viruses, including *Mimiviridae*, *Phycodnaviridae*, *Iridoviridae*, and *Poxviridae*, were also captured in the >0.45- μm -size fractions. While some of these larger viruses were small enough to pass through the 0.45- μm -pore-size filters, they may have been attached to cells, including nonhost cells, or particles.

We expect our observations of clear taxonomic differences in virus relative abundances in different size fractions to apply broadly to other environments. Our data show that examining only the smaller-size fraction can lead to underestimations of the abundance and diversity of certain virus groups such as virophages and cyanophages that could, in turn, lead to incorrect assumptions about their limited ecological importance in some environments. Furthermore, the fact that virus community profiles were vastly different when comparing viruses in the <0.45- μm - versus >0.45- μm -size fractions indicates that analysis of either size fraction alone provides only a partial perspective of environmental viruses. Conversely, viruses detected in both size fractions may be candidates for further exploration to discern active viral infections in the environment. Given our observation of unique ecological patterns associated with distinct size fractions, we recommend caution when interpreting environmental virus community assemblages and dynamics based on metagenomic data derived from different size fractions.

MATERIALS AND METHODS

Sampling sites and collection. Water samples were collected from three sites in Hamilton Harbour on 20 September 2019. Site 1 (43°17.253'N, 79°50.583'W) is located in the middle of the harbor at its deepest point, approximately at the location of the mid-harbor site from our 2015 sampling (14). Site 3 (43°16.904'N, 79°52.527'W) is closer to the shoreline and corresponds to the nearshore site from our sampling in 2015, and site 2 (43°17.135'N, 79°51.419'W) is located at the approximate midpoint between sites 1 and 3. Surface water samples were collected in 20-liter carboys using a bucket and were prefiltered through 100- μm -pore-size Nitex mesh to remove terrestrial plant debris and metazoan tissues, since this biomass could overwhelm the sequencing library. An RBRmaestro³ multichannel logger (RBR Ltd., Canada) was used to measure environmental physiochemical parameters, including temperature, conductivity, dissolved O₂, chlorophyll *a*, photosynthetically active radiation (PAR), salinity, and pH, and the Secchi depth at each site was also recorded.

Upon collection, water samples were covered and kept in the dark for immediate transportation and processing. After transportation to the lab, approximately 11 liters of each sample was pressure filtered (100 mm Hg maximum pressure) through 0.5- μm -nominal-pore-size GC50 glass fiber filters placed atop 0.45- μm -pore-size polyvinylidene fluoride (PVDF) filters (see Fig. S1 in the supplemental material). Following filtration, filters were stored at -20°C until DNA extraction. Filtrates, or water passing through the GC50 and PVDF filters, were treated with FeCl₃ to flocculate and precipitate viruses according to the protocol of John et al. (29), except 0.45- μm -pore-size PVDF filters were used to capture iron flocs after incubation. Filters with iron flocs were stored in the dark at 4°C until resuspension of viruses was performed according to the standard protocol (29). Resuspended viruses were transferred to polypropylene centrifuge tubes on top of a 1-ml sucrose cushion and pelleted via centrifugation at 182,000 $\times g$ for 3 h at 4°C. Supernatants were decanted, and 100 μl of Tris-EDTA (TE) buffer was added to each pellet, which was covered with Parafilm to soften at 4°C before resuspension and nucleic acid extraction.

To fulfill our goal of comparing virus communities in the <0.45- μm - and >0.45- μm -size fractions, we utilized two different capture methods to maximize nucleic acid recovery from these size fractions. For the >0.45- μm -size fractions, there was no additional concentration step beyond the filtration of water through the GC50 and PVDF filters. However, for the <0.45- μm -size fractions, comparable filter extraction methods are not possible. Filters with pore sizes small enough to capture viruses have very low porosity, allowing filtration of only small volumes; therefore, only minute amounts of DNA can be recovered from these filters, prohibiting effective library preparation for metagenomic sequencing. Therefore, for the <0.45- μm -size fractions, viruses were concentrated using FeCl₃ flocculation, a commonly used approach for concentrating free viruses that uses redox chemistry to flocculate and

TABLE 1 Summary of sample parameters throughout the processing pipeline

Sample	Concn (ng/ μ l)		No. of reads		No. of contigs	
	Initial DNA	Library	Pre-QC	Post-QC	Post-assembly (mean length [bp])	Annotated
<0.45 μ m						
Site 1	61.2	7.22	34,682,792	26,909,942	350,572 (699)	102,574
Site 2	82.7	6.53	33,438,900	25,831,282	329,713 (714)	97,968
Site 3	99.6	6.07	36,189,024	28,542,680	377,348 (745)	134,843
>0.45 μ m						
Site 1	43.8	4.24	35,222,034	25,937,694	290,644 (714)	164,583
Site 2	34.0	5.82	37,718,578	28,401,490	356,847 (756)	217,580
Site 3	45.8	5.27	40,863,560	28,521,306	192,849 (537)	45,468

subsequently liberate viruses in water samples (29). FeCl_3 flocculation was selected for its benefits over other widely used methods for concentrating aquatic viruses such as tangential flow filtration (TFF), which is expensive, time consuming, and produces highly variable results (29). Although using a flocculation approach was an essential aspect of this study that allowed us to compare virus communities in the <0.45- μ m- versus >0.45- μ m-size fractions, it is possible that this method concentrated certain types of viruses more effectively than others, thereby influencing estimates of virus diversity and relative abundance. However, such taxon-specific biases have not previously been identified with this viral recovery method.

DNA extraction and sequencing. For the >0.45- μ m-size fractions, one-quarter of each PVDF and GC50 filter was cut into pieces that were transferred to separate screw-cap centrifuge tubes. Community DNA was extracted separately from each filter using the DNeasy plant minikit (Qiagen, Germany) according to the manufacturer's protocol, except that bead beating instead of the TissueRuptor was used to disrupt samples. Approximately 0.25 g each of small (212 to 300 μ m) and large (425 to 600 μ m) glass beads was added to the centrifuge tubes with AP1 buffer and RNase A. The tubes were shaken in a FastPrep 96 bead beater (MP Biomedicals, USA) for 3 cycles of 90 s followed by 1 min on ice between cycles. Final DNA concentrations were measured using a Qubit dsDNA HS assay kit (Life Technologies, USA). Rather than combining equal volumes or amounts of DNA from each filter, we chose to pool the DNA proportionally to avoid biasing the "total" DNA sample with DNA extracted from only one filter. Ultimately, the different amounts of DNA recovered from each filter type were proportionally represented in the DNA sample that was sent for sequencing. For the <0.45- μ m-size fractions, virus pellets were resuspended by vortexing, and community DNA was extracted using the QIAamp viral RNA minikit (Qiagen, Germany) according to the manufacturer's protocol.

Final DNA concentrations were measured using a Qubit dsDNA HS assay kit (Life Technologies, USA), and DNA samples were sent for sequencing at the Microbial Genome Sequencing Centre (MiGS, Pittsburgh, PA). Library preparation was performed using a Nextera Flex DNA sample preparation kit (Illumina, USA) and sequenced from paired ends on the NextSeq 550 (2 by 150 bp) system (Illumina, USA).

Metagenome data processing. Demultiplexing and adapter removal were performed by MiGS using bcl2fastq. Raw reads were processed according to Palermo et al. (14) with minor alterations to account for the different read lengths. Briefly, quality control was performed using Sickel version 1.33 (61) using a quality cutoff value of 30 and read length cutoff value of 50. Reads were assembled using IDBA-UD version 1.1.3 (62) with k values from 20 to 100 in increments of 20 and a minimum k-mer count of 2. BLASTx searches against the December 2019 NCBI-nr database (downloaded from <https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz>) were performed using DIAMOND version 0.9.29 (63) with frameshift alignment and very sensitive modes activated. MEGAN6-LR version 6.14.2 (64) was used to annotate contigs using the November 2018 protein accession mapping file (downloaded from <https://software-ab.informatik.uni-tuebingen.de/download/megan6/welcome.html>) with long-read mode activated and bit score and E value cutoffs of 100 and 10^{-6} , respectively. Bowtie 2 version 2.3.5.1 (65) was used to map quality-controlled contigs back to the assembled contigs in very sensitive and end-to-end modes. SAMtools version 1.9 (66) was used to extract the mapping information for each contig. Table 1 summarizes the number of reads and contigs at each step in the processing pipeline for each sample, while Table S1 provides statistical information for the subset of contigs annotated as viral in each sample. Figure S2 shows the length distribution of contigs annotated as viral in the samples from each size fraction.

Metagenome data analysis. Data analysis methods were according to those previously described in detail by Palermo et al. (14). Briefly, contig relative abundances were calculated by mapping quality-controlled reads back to assembled contigs and dividing the number of mapped reads by the length of the contig to normalize by contig length. These values were summed for all contigs in a sample and divided by 1,000 to generate a normalized sum that was used to account for differences in sequencing depth between samples. All virus contig assignments were categorized as one of the following based on NCBI taxonomic classifications: *Caudovirales*, virophages (*Lavidaviridae*, *Mimiviridae*, *Phycodnaviridae*, *Iridoviridae*, *Poxviridae*, other bacterial viruses, other dsDNA viruses, ssDNA viruses, and unclassified

viruses. Taxonomic assignments for several contigs were less specific than information provided in the literature or NCBI's taxonomy browser; therefore, these contigs were manually reassigned to more-specific groups (Table S2). Additionally, there were some contigs originally annotated as *Phycodnaviridae* that are now considered to be more closely related to *Mimiviridae* and are placed within the proposed subfamily "*Mesomimivirinae*" (17, 67). These contigs were manually reassigned to the *Mimiviridae* category. *Marseilleviridae* and *Tectiviridae* were present at low relative abundances (<0.01% and <0.02%, respectively) and were therefore grouped together with the "other dsDNA viruses" (Table S2).

To assess the potential impact of *Caudovirales* on the prokaryotic community, *Caudovirales* contigs with inferred hosts that could be inferred from database information were extracted and analyzed separately. Host inferences were determined based on the name of the annotated virus. For example, a virus named "*Synechococcus* phage" was presumed to infect *Synechococcus*, whereas a virus named "Mediterranean phage" was presumed to infect an unknown host and was not included in the analysis of *Caudovirales* with inferred hosts.

Various packages in RStudio were used for data analysis and visualization as follows: virus relative abundances were presented as bar charts and bubble plots using the ggplot2 package (68), unique contig annotations were compared between the size fractions using Venn diagrams generated using the VennDiagram package (69), and Bray-Curtis dissimilarity analyses with the unweighted pair group with arithmetic mean (UPGMA) clustering were generated using the vegan and dendextend packages (70, 71).

Data availability. DNA sequences generated in this study were deposited in the NCBI Sequence Read Archive under the BioProject accession number [PRJNA670357](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA670357).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.3 MB.

ACKNOWLEDGMENTS

DNA extraction and metagenomic sequencing was supported by an NSERC Discovery Grant (RGPIN-2016-06022) awarded to S.M.S.

We thank Bailey McMeans for the use of her sampling boat and RBRmaestro3 multichannel logger. We also thank Rose Mastin Wood for assistance with sampling and sample processing.

Conceptualization, C.N.P., D.W.S., and S.M.S.; methodology, C.N.P., D.W.S., and S.M.S.; investigation, C.N.P., D.W.S., and S.M.S.; data curation, C.N.P. and S.M.S.; formal analysis, C.N.P. and S.M.S.; validation, C.N.P. and S.M.S.; visualization, C.N.P. and S.M.S.; writing—original draft preparation, C.N.P. and S.M.S.; writing—review and editing, C.N.P., D.W.S., and S.M.S.; supervision, S.M.S.; project administration, S.M.S.; resources, S.M.S.; funding acquisition, S.M.S.

REFERENCES

- Thingstad TF, Lignell R. 1997. Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat Microb Ecol* 13:19–27. <https://doi.org/10.3354/ame013019>.
- Winter C, Bouvier T, Weinbauer MG, Thingstad TF. 2010. Trade-offs between competition and defense specialists among unicellular planktonic organisms: the "killing the winner" hypothesis revisited. *Microbiol Mol Biol Rev* 74:42–57. <https://doi.org/10.1128/MMBR.00034-09>.
- Zimmerman AE, Howard-Varona C, Needham DM, John SG, Worden AZ, Sullivan MB, Waldbauer JR, Coleman ML. 2020. Metabolic and biogeochemical consequences of viral infection in aquatic ecosystems. *Nat Rev Microbiol* 18:21–34. <https://doi.org/10.1038/s41579-019-0270-x>.
- Koonin EV. 2006. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biol Direct* 1:22. <https://doi.org/10.1186/1745-6150-1-22>.
- Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient virus world and evolution of cells. *Biol Direct* 1:29. <https://doi.org/10.1186/1745-6150-1-29>.
- Roux S, Brum JR, Dutilh BE, Sunagawa S, Duhaime MB, Loy A, Poulos BT, Solonenko N, Lara E, Poulain J, Pesant S, Kandels-Lewis S, Dimier C, Picheral M, Searson S, Cruaud C, Alberti A, Duarte CM, Gasol JM, Vaque D, Tara Oceans Coordinators, Bork P, Acinas SG, Wincker P, Sullivan MB. 2016. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* 537:689–693. <https://doi.org/10.1038/nature19366>.
- Zhou JL, Zhang WJ, Yan SL, Xiao JZ, Zhang YY, Li BL, Pan YJ, Wang YJ. 2013. Diversity of virophages in metagenomic data sets. *J Virol* 87:4225–4236. <https://doi.org/10.1128/JVI.03398-12>.
- Brum JR, Sullivan MB. 2015. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat Rev Microbiol* 13:147–159. <https://doi.org/10.1038/nrmicro3404>.
- Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, de Vargas C, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans Coordinators, Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Patterns and ecological drivers of ocean viral communities. *Science* 348:1261498. <https://doi.org/10.1126/science.1261498>.
- Paez-Espino D, Elie-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, Rubin E, Ivanova NN, Kyrpides NC. 2016. Uncovering Earth's virome. *Nature* 536:425–430. <https://doi.org/10.1038/nature19094>.
- Roux S, Hallam SJ, Woyke T, Sullivan MB. 2015. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *Elife* 4:e08490. <https://doi.org/10.7554/eLife.08490>.
- Paez-Espino D, Zhou JL, Roux S, Nayfach S, Pavlopoulos GA, Schulz F, McMahon KD, Walsh D, Woyke T, Ivanova NN, Elie-Fadrosh EA, Tringe SG, Kyrpides NC. 2019. Diversity, evolution, and classification of virophages uncovered through global metagenomics. *Microbiome* 7:157. <https://doi.org/10.1186/s40168-019-0768-5>.
- Schulz F, Roux S, Paez-Espino D, Jungbluth S, Walsh DA, Denef VJ, McMahon KD, Konstantinidis KT, Elie-Fadrosh EA, Kyrpides NC, Woyke T. 2020. Giant virus diversity and host interactions through global metagenomics. *Nature* 578:432–436. <https://doi.org/10.1038/s41586-020-1957-x>.

14. Palermo CN, Fulthorpe RR, Saati R, Short SM. 2019. Metagenomic analysis of virus diversity and relative abundance in a eutrophic freshwater harbour. *Viruses* 11:792. <https://doi.org/10.3390/v11090792>.
15. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* 99:14250–14255. <https://doi.org/10.1073/pnas.202488399>.
16. Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F. 2004. Diversity and population structure of a near-shore marine-sediment viral community. *Proc Biol Sci* 271:565–574. <https://doi.org/10.1098/rspb.2003.2628>.
17. Claverie J-M, Abergel C. 2018. *Mimiviridae*: an expanding family of highly diverse large dsDNA viruses infecting a wide phylogenetic range of aquatic eukaryotes. *Viruses* 10:506. <https://doi.org/10.3390/v10090506>.
18. Vibin J, Chamings A, Collier F, Klaassen M, Nelson TM, Alexandersen S. 2018. Metagenomics detection and characterisation of viruses in faecal samples from Australian wild birds. *Sci Rep* 8:8686. <https://doi.org/10.1038/s41598-018-26851-1>.
19. Deng L, Silins R, Castro-Mejia JL, Kot W, Jessen L, Thorsen J, Shah S, Stokholm J, Bisgaard H, Moineau S, Nielsen DS. 2019. A protocol for extraction of infective viromes suitable for metagenomics sequencing from low volume fecal samples. *Viruses* 11:667. <https://doi.org/10.3390/v11070667>.
20. Temmam S, Monteil-Bouchard S, Robert C, Pascalis H, Michelle C, Jardot P, Charrel R, Raoult D, Desnues C. 2015. Host-associated metagenomics: a guide to generating infectious RNA viromes. *PLoS One* 10:e0139810. <https://doi.org/10.1371/journal.pone.0139810>.
21. Bekliz M, Brandani J, Bourquin M, Battin TJ, Peter H. 2019. Benchmarking protocols for the metagenomic analysis of stream biofilm viromes. *PeerJ* 7:e8187. <https://doi.org/10.7717/peerj.8187>.
22. Goller PC, Haro-Moreno JM, Rodriguez-Valera F, Loessner MJ, Gomez-Sanz E. 2020. Uncovering a hidden diversity: optimized protocols for the extraction of dsDNA bacteriophages from soil. *Microbiome* 8:17. <https://doi.org/10.1186/s40168-020-0795-2>.
23. Conceicao-Neto N, Zeller M, Lefrere H, De Bruyn P, Beller L, Deboutte W, Yinda CK, Lavigne R, Maes P, Van Ranst M, Heylen E, Matthijssens J. 2015. Modular approach to customise sample preparation procedures for viral metagenomics: a reproducible protocol for virome analysis. *Sci Rep* 5:16532. <https://doi.org/10.1038/srep16532>.
24. Nakai R. 2020. Size matters: ultra-small and filterable microorganisms in the environment. *Microb Environ* 35:ME20025. <https://doi.org/10.1264/jsme2.ME20025>.
25. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. 2009. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 4:470–483. <https://doi.org/10.1038/nprot.2009.10>.
26. Sun GW, Xiao JZ, Wang HM, Gong CW, Pan YJ, Yan SL, Wang YJ. 2014. Efficient purification and concentration of viruses from a large body of high turbidity seawater. *MethodsX* 1:197–206. <https://doi.org/10.1016/j.mex.2014.09.001>.
27. Roux S, Krupovic M, Debroas D, Forterre P, Enault F. 2013. Assessment of viral community functional potential from viral metagenomes may be hampered by contamination with cellular sequences. *Open Biol* 3:130160. <https://doi.org/10.1098/rsob.130160>.
28. Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. 2010. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* 18:11–19. <https://doi.org/10.1016/j.tim.2009.11.003>.
29. John SG, Mendez CB, Deng L, Poulos B, Kauffman AKM, Kern S, Brum J, Polz MF, Boyle EA, Sullivan MB. 2011. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* 3:195–202. <https://doi.org/10.1111/j.1758-2229.2010.00208.x>.
30. Palermo CN, Shea DW, Short SM. 4 November 2020. Comparing the diversity and relative abundance of free and particle-associated aquatic viruses. *bioRxiv* <https://doi.org/10.1101/2020.11.03.367664>.
31. Roux S, Chan LK, Egan R, Malmstrom RR, McMahon KD, Sullivan MB. 2017. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nat Commun* 8:858. <https://doi.org/10.1038/s41467-017-01086-2>.
32. Hendrix RW, Smith MCM, Burns RN, Ford ME, Hatfull GF. 1999. Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. *Proc Natl Acad Sci U S A* 96:2192–2197. <https://doi.org/10.1073/pnas.96.5.2192>.
33. Bannert N, Kurth R. 2004. Retroelements and the human genome: new perspectives on an old relation. *Proc Natl Acad Sci U S A* 101:14572–14579. <https://doi.org/10.1073/pnas.0404838101>.
34. Forterre P. 2013. The virocell concept and environmental microbiology. *ISME J* 7:233–236. <https://doi.org/10.1038/ismej.2012.110>.
35. Potapov SA, Tikhonova IV, Krasnopeev AY, Kabilov MR, Tupikin AE, Chebunina NS, Zhuchenko NA, Belykh OI. 2019. Metagenomic analysis of viroplankton from the pelagic zone of Lake Baikal. *Viruses* 11:991. <https://doi.org/10.3390/v11110991>.
36. Skvortsov T, de Leeuwe C, Quinn JP, McGrath JW, Allen CC, McElarney Y, Watson C, Arkhipova K, Lavigne R, Kulakov LA. 2016. Metagenomic characterisation of the viral community of Lough Neagh, the largest freshwater lake in Ireland. *PLoS One* 11:e0150361. <https://doi.org/10.1371/journal.pone.0150361>.
37. Green JC, Rahman F, Saxton MA, Williamson KE. 2015. Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern Virginia, USA. *Aquat Microb Ecol* 75:117–128. <https://doi.org/10.3354/ame01752>.
38. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. 2012. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* 7:e33641. <https://doi.org/10.1371/journal.pone.0033641>.
39. Mohiuddin M, Schellhorn HE. 2015. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Front Microbiol* 6:960. <https://doi.org/10.3389/fmicb.2015.00960>.
40. Ge X, Wu Y, Wang M, Wang J, Wu L, Yang X, Zhang Y, Shi Z. 2013. Viral metagenomics analysis of planktonic viruses in East Lake, Wuhan, China. *Virol Sin* 28:280–290. <https://doi.org/10.1007/s12250-013-3365-y>.
41. Lopez-Bueno A, Tamames J, Velazquez D, Moya A, Quesada A, Alcamí A. 2009. High diversity of the viral community from an Antarctic lake. *Science* 326:858–861. <https://doi.org/10.1126/science.1179287>.
42. Butina TV, Bukin YS, Krasnopeev AS, Belykh OI, Tupikin AE, Kabilov MR, Sakirko MV, Belikov SI. 2019. Estimate of the diversity of viral and bacterial assemblage in the coastal water of Lake Baikal. *FEMS Microbiol Lett* 366:fnz094. <https://doi.org/10.1093/femsle/fnz094>.
43. Watkins SC, Kuehnle N, Ruggeri CA, Malki K, Bruder K, Elayyan J, Damisch K, Vahora N, O'Malley P, Ruggles-Sage B, Romer Z, Putonti C. 2016. Assessment of a metaviromic dataset generated from nearshore Lake Michigan. *Mar Freshw Res* 67:1700–1708. <https://doi.org/10.1071/MF15172>.
44. Munawar M, Fitzpatrick M, Niblock H, Kling H, Rozon R, Lorimer J. 2017. Phytoplankton ecology of a culturally eutrophic embayment: Hamilton Harbour, Lake Ontario. *Aquat Ecosyst Health Manag* 20:201–213. <https://doi.org/10.1080/14634988.2017.1307678>.
45. Munawar M, Fitzpatrick M. 2007. An integrated assessment of the microbial and planktonic communities of Hamilton Harbour. *Can Tech Rep Fish Aquat Sci* 2729:43–63.
46. Hotto AM, Satchwell MF, Boyer GL. 2007. Molecular characterization of potential microcystin-producing cyanobacteria in Lake Ontario embayments and nearshore waters. *Appl Environ Microbiol* 73:4570–4578. <https://doi.org/10.1128/AEM.00318-07>.
47. Harke MJ, Steffen MM, Gobler CJ, Otten TG, Wilhelm SW, Wood SA, Paerl HW. 2016. A review of the global ecology, genomics, and biogeography of the toxic cyanobacterium, *Microcystis* spp. *Harmful Algae* 54:4–20. <https://doi.org/10.1016/j.hal.2015.12.007>.
48. Wilhelm SW, Bullerjahn GS, McKay RML. 2020. The complicated and confusing ecology of *Microcystis* blooms. *mBio* 11:e00529–20. <https://doi.org/10.1128/mBio.00529-20>.
49. Yanagida F, Chen YS, Yasaki M. 2007. Isolation and characterization of lactic acid bacteria from lakes. *J Basic Microbiol* 47:184–190. <https://doi.org/10.1002/jobm.200610237>.
50. Kang YH, Kang SK, Park CS, Joo JH, Lee JW, Han MS. 2016. Use of lactic acid bacteria as a biological agent against the cyanobacterium *Anabaena flos-aquae*. *J Appl Phycol* 28:1747–1757. <https://doi.org/10.1007/s10811-015-0701-7>.
51. Lapidus A, Clum A, LaButti K, Kaluzhnyaya MG, Lim S, Beck DAC, del Rio TG, Nolan M, Mavromatis K, Huntemann M, Lucas S, Lidstrom ME, Ivanova N, Chistoserdova L. 2011. Genomes of three methylotrophs from a single niche reveal the genetic and metabolic divergence of the *Methylophilaceae*. *J Bacteriol* 193:3757–3764. <https://doi.org/10.1128/JB.00404-11>.
52. Paver SF, Newton RJ, Coleman ML. 2020. Microbial communities of the Laurentian Great Lakes reflect connectivity and local biogeochemistry. *Environ Microbiol* 22:433–446. <https://doi.org/10.1111/1462-2920.14862>.
53. Chistoserdova L. 2011. Methylotrophy in a lake: from metagenomics to single-organism physiology. *Appl Environ Microbiol* 77:4705–4711. <https://doi.org/10.1128/AEM.00314-11>.
54. Kalyuzhnyaya MG, Martens-Habbena W, Wang TS, Hackett M, Stolyar SM, Stahl DA, Lidstrom ME, Chistoserdova L. 2009. *Methylophilaceae* link

- methanol oxidation to denitrification in freshwater lake sediment as suggested by stable isotope probing and pure culture analysis. *Environ Microbiol Rep* 1:385–392. <https://doi.org/10.1111/j.1758-2229.2009.00046.x>.
55. Coutinho FH, Cabello-Yeves PJ, Gonzalez-Serrano R, Rosselli R, Lopez-Perez M, Zenskaya TI, Zakharenko AS, Ivanov VG, Rodriguez-Valera F. 2020. New viral biogeochemical roles revealed through metagenomic analysis of Lake Baikal. *Microbiome* 8:163. <https://doi.org/10.1186/s40168-020-00936-4>.
 56. Moon K, Kang I, Kim S, Kim SJ, Cho JC. 2017. Genome characteristics and environmental distribution of the first phage that infects the LD28 clade, a freshwater methylotrophic bacterial group. *Environ Microbiol* 19:4714–4727. <https://doi.org/10.1111/1462-2920.13936>.
 57. Cho BC, Hardies SC, Jang GI, Hwang CY. 2018. Complete genome of streamlined marine actinobacterium *Pontimonas salivibrio* strain CL-TW6^T adapted to coastal planktonic lifestyle. *BMC Genomics* 19:625. <https://doi.org/10.1186/s12864-018-5019-9>.
 58. Breitbart M, Rohwer F. 2005. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* 13:278–284. <https://doi.org/10.1016/j.tim.2005.04.003>.
 59. Fischer MG, Hackl T. 2016. Host genome integration and giant virus-induced reactivation of the virophage mavirus. *Nature* 540:288–291. <https://doi.org/10.1038/nature20593>.
 60. Desnues C, Raoult D. 2010. Inside the lifestyle of the virophage. *Intervirology* 53:293–303. <https://doi.org/10.1159/000312914>.
 61. Joshi NA, Fass JN. 2011. Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files (version 1.33). <https://github.com/najoshi/sickle>.
 62. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>.
 63. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 12:59–63. <https://doi.org/10.1038/nmeth.3176>.
 64. Huson DH, Albrecht B, Bağcı C, Bessarab I, Górski A, Jolic D, Williams RBH. 2018. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol Direct* 13:6. <https://doi.org/10.1186/s13062-018-0208-7>.
 65. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>.
 66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>.
 67. Blanc G, Gallot-Lavallee L, Maumus F. 2015. Provirophages in the *Bigeloviella* genome bear testimony to past encounters with giant viruses. *Proc Natl Acad Sci U S A* 112:E5318–E5326. <https://doi.org/10.1073/pnas.1506469112>.
 68. Wickham H. 2009. ggplot2: elegant graphics for data analysis. Springer, New York, NY.
 69. Chen H, Boutros PC. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* 12:35. <https://doi.org/10.1186/1471-2105-12-35>.
 70. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2017. vegan: Community Ecology Package. R package version 2.4–5. <https://CRAN.R-project.org/package=vegan>.
 71. Galili T. 2015. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 31:3718–3720. <https://doi.org/10.1093/bioinformatics/btv428>.