Metagenomic analysis of virus diversity and relative abundance in a eutrophic freshwater harbour

Authors: Christine N. Palermo[1], Roberta R. Fulthorpe[2], Rosemary Saati[2], and Steven M. Short[1]

[1]University of Toronto Mississauga

3359 Mississauga Road, Mississauga, Ontario, Canada

[2]University of Toronto Scarborough

1265 Military Trail, Scarborough, Ontario, Canada

Christine N. Palermo: christine.palermo@mail.utoronto.ca

Roberta R. Fulthorpe: fulthorpe@utsc.utoronto.ca

Rosemary Saati: rosemary.saati@mail.utoronto.ca

Steven M. Short (corresponding author): steven.short@utoronto.ca

Keywords: metagenomics, viral ecology, freshwater virology, microbial communities, virophages

ABSTRACT

Aquatic viruses have been extensively studied over the past decade, yet fundamental aspects of freshwater virus communities remain poorly described. Our goal was to characterize particle-associated virus communities seasonally and spatially in a freshwater harbour. Community DNA was extracted from water samples and sequenced on an Illumina HiSeq platform. Assembled contigs were annotated as belonging to the virus families *Caudovirales*, *Mimiviridae*, *Phycodnaviridae*, and virophages (*Lavidaviridae*), or to other groups of undefined viruses. Diverse *Mimiviridae* contigs were detected in the samples, but the two sites contained distinct *Mimiviridae* communities. Virophages were often the most abundant group, and discrete virophage taxa were remarkably stable across sites and dates despite fluctuations in *Mimiviridae* community composition. *Caudovirales* were present at low abundances in most samples, contrasting other studies of freshwater environments. Similarly, *Phycodnaviridae* abundances were surprisingly low in all samples despite the harbour's capacity to support high algal biomass during the summer and autumn months, suggesting that *Mimiviridae* are the dominant algae-infecting viruses in this system. Overall, our findings provided insights into freshwater virus community assemblages by expanding the documented diversity of freshwater virus communities, highlighting the potential ecological importance of virophages, and revealing distinct communities over small spatial scales.

INTRODUCTION

Viruses can modify and control the structure and function of ecosystems and in turn, influence global biogeochemical cycles and the evolution of organisms [1]. Historically, use of traditional culture-based methods to study environmental viruses led to underestimations of their abundance and ecological importance [2], whereas more contemporary molecular methods allowed analyses of environmental microbial communities without the constraints of cultivation. However, unlike the prokaryotes and eukaryotes they infect, viruses do not share universally conserved genes that can be readily targeted to survey entire communities. Thus, viral ecology is a field for which major advances can be realized through shotgun metagenomic sequencing [3]. Nonetheless, despite intensive efforts over the past couple of decades, comprehensive knowledge of virus community diversity and dynamics remains elusive for most natural settings.

In the first viral metagenomics study, over 65% of sequences recovered from surface seawater samples were not significantly similar to any sequence in existing databases, highlighting the lack of knowledge of environmental viruses [4]. Since then, viral sequence databases have expanded dramatically, in large part due to several large-scale ocean sampling expeditions that have included viral community analysis (as reviewed in [5]). These sampling expeditions include the Tara Oceans Expedition, the Malaspina Circumnavigation Expedition, Pacific Ocean Virome (POV), the San Pedro Ocean Time-series (SPOT), and the Bermuda Atlantic Time-series Study (BATS). BATS tracked viral abundance in the Sargasso Sea over a decade, while the SPOT studies measured temporal variation in virus communities and their hosts. The POV was established with data from transects spanning from coastal waters to the open ocean to document

spatial changes in microbial and virus communities, whereas the Tara Oceans and Malaspina Circumnavigation Expeditions were designed to gather baseline global oceanic biodiversity data.

The Tara Oceans Expedition was conducted from 2009-2013 with the aim of globally sampling a wide range of organismal and functional diversity in the surface oceans, while the Malaspina Circumnavigation Expedition sailed from December 2010 to July 2011 with a focus on deep ocean microbiology. The Tara Oceans Expedition resulted in several important discoveries related to diverse groups of marine planktonic taxa, including viruses (e.g. [6-10]). Importantly, this sampling expedition provided data supporting previous observations of high local diversity but limited global diversity that led to the conception of a 'seed-bank' model of virus diversity [11]. The Tara Oceans survey revealed that virus community composition was strongly impacted by temperature and oxygen concentrations on local scales due to these factors' influence on their hosts, while on larger scales ocean currents were responsible for transporting and mixing a virus 'seed-bank' [6]. Furthermore, using 17 viromes generated from the expedition, the abundance and diversity of nucleo-cytoplastic large DNA viruses (NCLDVs) were mapped revealing that there were approximately 10^4-10^5 viruses per ml in the photic zone and that the so-called 'Megavirales' and the *Phycodnaviridae* were the most common NCLDVs in the epipelagic oceans [12]. Complementing the Tara Oceans Expedition, data stemming from the Malaspina Circumnavigation Expedition demonstrated that viruses have higher turnover rates in the deep ocean compared surface waters, and they play important roles in DOC production and nutrient release, especially in the bathypelagic [13]. By combining viral sequences from the Tara Oceans and the Malaspina Circumnavigation Expeditions, numerous virus genomes have been assembled, expanding viral sequence databases more than three-fold [14]. Thus, these

83 expeditions have vastly improved our understanding of marine viral ecology and have
84 highlighted the global importance of virus activity.
85
86 Though extensive surveys of marine virus communities have been conducted, relatively little is
87 known about fundamental aspects of freshwater virus ecology, such as their distribution in the
88 environment. Despite their underrepresentation in databases, research has demonstrated that
89 freshwater virus communities contain novel viruses and are distinct from other aquatic virus
90 communities [15-17]. Metagenomics has been used to study virus communities in natural
91 freshwater lakes from the Arctic [18], Canada [19], USA [20,21], Ireland [22], France [17],
92 China [23], and Antarctica [24,25]. With respect to virus communities in eutrophic lakes, studies
93 by Green et al. [20], Skvortsov et al. [22], and Ge et al. [23] revealed virus communities
94 dominated by *Caudovirales* in the epilimnion of eutrophic lakes in China, USA, and Ireland,
95 respectively, but other dsDNA viruses, unclassified bacteriophages, and ssDNA viruses were
96 also detected albeit at lower abundances. Roux et al. [26] focused on the virophage and NCLDV
97 communities in a eutrophic freshwater lake in USA, and observed highly dynamic virophage
98 communities lacking any apparent annual or seasonal patterns of abundance. Although viral
99 sequence databases have expanded rapidly alongside knowledge of marine virus ecology,
100 documenting virus diversity and activity in disparate freshwater environments remains an
101 exciting and untapped area for research.
102
103 The goal of the research described here was to characterize the virus community in Hamilton
104 Harbour, a eutrophic freshwater embayment of Lake Ontario. The harbour is located at the
105 western end of Lake Ontario and has an area of 21.5-km^2 and an average depth of 13 m. It is

106 separated from Lake Ontario by a naturally occurring sandbar and the Burlington Shipping Canal
107 [27], and is the largest Canadian port in the Great Lakes [28]. The area surrounding the harbour
108 has a long history of industrial activity, resulting in a highly polluted harbour containing heavy
109 metals and many other hazardous contaminants [29,30]. Moreover, inputs from wastewater
110 treatment plants, stormwater and sewage overflow, and agricultural and urban runoff have led to
111 high nutrient concentrations, especially phosphorus. Thus, Hamilton Harbour is a eutrophic
112 system that experiences seasonal blooms of cyanobacteria and algae, poor water clarity, and
113 depleted hypolimnetic oxygen. As a result of all these perturbations, the harbour was designated
114 an 'Area of Concern' in the amended 1987 USA-Canada Great Lakes Water Quality Agreement
115 (GLWQA). Despite its designation over 30 years ago and ongoing remediation efforts, Hamilton
116 Harbour remains one of the most impaired sites in the Canadian Great Lakes [31].
117
118 While Hamilton Harbour is in general an extensively studied system, the microbial community
119 has only been examined using microscopic techniques to investigate microbial diversity and
120 abundance. To our knowledge, there are no published studies characterizing the microbial
121 community in Hamilton Harbour using metagenomics, nor are there any studies of the virus
122 community in Hamilton Harbour. The metagenomic dataset used in this study was originally
123 generated to study Hamilton Harbour bacteria and eukaryotic plankton, but was leveraged herein
124 to provide an initial assessment of the cell- and particle-associated virus community and to
125 determine if seasonal and spatial patterns of diversity and abundance could be discerned in this
126 unique freshwater environment. More research is required to better characterize freshwater virus
127 diversity, community structures, and patterns of abundance, and the factors that drive these

128 phenomena. This research aims to address these knowledge gaps by providing a detailed view of
129 the virus community in Hamilton Harbour.
130
131 MATERIALS AND METHODS
132 *Sampling Sites and Collection*
133 Water samples were collected from two long-term Environment Canada monitoring sites in
134 Hamilton Harbour (referred to as stations 1001 and 9031 in previous publications). One site
135 (station 1001) is located at the deepest and most central part of the harbour (43°17'17.0"N
136 79°50'23.0"W) with a water depth of 24 m and a 1.2 km distance from the shoreline. The other
137 site (station 9031) is located less than 0.5 km from the shoreline (43°16'50.0"N 79°52'32.0"W),
138 with a water depth of 12 m. This "nearshore" site is influenced by effluents from the Cootes
139 Paradise watershed on the west end of the harbour, while the "mid-harbour" site is closer to the
140 Burlington Shipping Canal and Lake Ontario. In 2015, water samples for metagenomic analyses
141 were collected on July 30th, August 13th and 27th, and September 10th and 24th, from both sites at
142 approximately 1 m below the surface using a Van Dorn bottle sampler; in total, 10 samples were
143 collected, 5 from each of the nearshore and mid-harbour sites. 500 ml water samples were
144 filtered through 0.22 μm pore-size Sterivex capsule filters (EMD Millipore, USA), and the filters
145 were sealed and stored at -80°C until further analysis. For each sample collected, physiochemical
146 parameters including pH, dissolved oxygen, temperature, redox potential, and chlorophyll a were
147 measured *in situ* using a YSI 58 (Xylem Inc., USA). Secchi depth was also measured with each
148 sample.
149
150 *DNA Extraction and Sequencing*

151 To extract community DNA from each sample, biomass was recovered from the filters by adding
152 2 ml of molecular grade nuclease free water to each filter and vortexing for 5 minutes. The
153 resuspended material was transferred to a 1.5 ml microcentrifuge tube under sterile conditions
154 and was centrifuged at 6,000 x g for a total of 15 minutes. DNA was extracted from the pelleted
155 material using a FastDNA SPIN Kit (MP BIO, USA), beginning with the addition of 500 µl of
156 CLS-Y solution. The manufacturer's protocol was followed except the wash step was repeated
157 three times to maximize removal of environmental contaminants. Following extraction, DNA
158 concentrations were estimated using a NanoDrop ND-1000 UV-Vis Spectrophotometer
159 (NanoDrop Technologies, USA), and were standardized to 1.5 µg of DNA per sample before
160 being sent for library preparation and shotgun metagenome sequencing by MR. DNA (Molecular
161 Research LP, USA).
162
163 Sample DNA libraries were prepared using a Nextera DNA Sample Preparation Kit (Illumina,
164 USA), following the manufacturer's protocol. DNA concentrations were measured using the
165 Qubit dsDNA HS Assay Kit (Life Technologies, USA) before and after library preparation
166 (Table 1). Prior to library preparation, samples were diluted to achieve the recommended
167 concentration of 2.5 ng μl^{-1}. For the August 27 nearshore and mid-harbour sites, achieving a
168 concentration of 2.5 ng μl^{-1} was not possible, so the maximum volume (20 µl) of each sample
169 was used for these libraries. Average library size was determined using an Agilent 2100
170 Bioanalyzer (Agilent Technologies, USA). Each library was clustered using a cBot System
171 (Illumina, USA), and was sequenced from paired ends using 500 cycles with the HiSeq 2500 (2 x
172 250 bp) system (Illumina, USA).
173

174 **Table 1:** Initial DNA concentration, final library concentration, and average fragment size in the
175 library for each of the 10 samples in this study.

Sample	Initial DNA Concentration (ng/µl)	Library Concentration (ng/µl)	Average Library Size (bp)
July 30 – Nearshore	5.30	16.2	985
July 30 – Mid-harbour	2.84	18.3	1070
August 13 – Nearshore	2.76	17.5	1070
August 13 – Mid-harbour	3.40	14.5	1000
August 27 – Nearshore	2.28*	15.0	950
August 27 – Mid-harbour	1.42*	8.94	615
September 10 – Nearshore	9.58	12.0	1000
September 10 – Mid-harbour	10.6	14.1	1060
September 24 – Nearshore	9.10	15.7	1030
September 24 – Mid-harbour	2.68	17.7	900

176 * denotes that samples could not be adjusted to 2.5 ng/µL prior to library preparation.

177

178 *Metagenome Data Processing*

179 After sequencing and base-calling, adapters and barcodes were removed by MR DNA
180 (Molecular Research LP, USA). Read quality parameters were verified using FastQC version
181 0.11.5 [32] prior to quality control and once again prior to assembly. Quality control was
182 performed using a sliding window method with the program Sickle version 1.33 [33] using a
183 quality score cut-off value of 30. All reads shorter than 50 bp were removed prior to assembly.
184 Reads were assembled using IDBA-UD version 1.1.3 [34] with alterations to the source code to
185 accommodate longer read lengths and higher maximum k values (as in [22]). The de Bruijn

graph-based assembly was performed using k values from 20 to 200 in increments of 20. A minimum k-mer count of 1 was used in the assembly to maximize assembly of the low coverage reads. Contig alignment was achieved with BLASTx against the April 2018 NCBI-nr database downloaded from ftp://ftp.ncbi.nih.gov/blast/db/FASTA/nr.gz. DIAMOND version 0.9.19 [35] was used for alignment with frameshift alignment and very sensitive modes activated. MEGAN6-LR version 6.11.4 [36] was used to annotate contigs using the Lowest Common Ancestor (LCA) algorithm in long read mode with a bit score cut-off value of 100 and a 10^{-6} e-value cut-off. The March 2018 MEGAN protein accession mapping file was downloaded from http://ab.inf.uni-tuebingen.de/data/software/megan6/download/welcome.html. Quality controlled reads were mapped back to assembled contigs using Bowtie 2 [37] in very sensitive mode, and mapping information for each contig was extracted using SAMtools [38]. Table 2 summarizes the number of reads and contigs at each step in the pipeline for each sample.

Table 2: Summary of the number of reads and contigs at each step in the data processing pipeline

Sample	Reads pre-QC	Reads post-QC	Contigs post-assembly (Mean length)	Contigs assigned	Virus contigs assigned	Percent of virus contigs
July 30 – Nearshore	13,403,832	11,608,892	480,233 (693)	212,156	349	0.16
July 30 – Mid-harbour	13,206,316	11,246,470	433,927 (824)	256,277	397	0.15
August 13 – Nearshore	12,758,364	11,040,814	398,774 (741)	196,128	488	0.25
August 13 – Mid-harbour	13,104,992	11,414,576	416,468 (821)	273,693	662	0.24
August 27 – Nearshore	15,685,750	13,953,600	374,464 (662)	128,233	484	0.38

August 27 – Mid-harbour	16,101,196	14,751,904	233,868 (834)	183,817	232	0.13
September 10 – Nearshore	16,228,148	14,257,158	418,490 (609)	110,567	463	0.42
September 10 – Mid-harbour	15,411,142	13,561,346	443,692 (568)	64,142	604	0.94
September 24 – Nearshore	14,689,040	12,764,018	367,778 (629)	112,144	372	0.33
September 24 – Mid-harbour	13,151,390	11,228,970	296,813 (777)	185,913	104	0.06

Metagenome Data Analysis

Once taxonomic assignments were applied to contigs, their relative abundances were estimated after normalizing by contig length and sequencing depth per sample. The number of reads that mapped to each contig was divided by the contig length in kbp. These values were summed for all contigs in a sample and were divided by a normalizing factor of 1000. Then, the number of mapped reads per kbp was divided by the normalized sum to achieve a relative abundance count that could be compared to other groups within and between samples. All virus contig assignments were sorted into one of the following categories based on the NCBI taxonomic classifications: *Caudovirales, Mimiviridae, Phycodnaviridae,* virophages (*Lavidaviridae*), *Iridoviridae*, *Poxviridae*, other dsDNA viruses, ssDNA viruses, unclassified bacterial viruses, and unclassified viruses. Because the assigned taxonomic classifications for some contigs in the "unclassified bacterial viruses" and "unclassified viruses" categories were less specific than information provided in the literature within which they were originally reported, some contigs in these categories were manually curated and were assigned to more specific groups based on published information (Table S1).

As well as assigning some contigs to more specific categories, some of the more specific assignments in the dsDNA viruses group were re-assigned to the "Other dsDNA viruses" group if they were observed in less than half of the samples and at <0.5% abundance. For example, contigs annotated as *Marseilleviridae* were only observed in 2 of the 10 samples at < 0.2 % abundance. Because there was not enough representation to making meaningful comparisons, the *Marseilleviridae* were grouped together with the "Other dsDNA viruses". As a final note, in one sample a single contig was annotated as an RNA virus, and this contig was re-assigned to the "Unclassified viruses" category (Table S1).

There is evidence that some viruses previously and tentatively considered phycodnaviruses (i.e., members of the *Phycodnaviridae* family) are more closely related to *Mimiviridae* than *Phycodnaviridae*, and in fact form a separate subfamily 'Mesomimivirinae' within the *Mimiviridae* [39]. Though not formally recognized by the International Committee for Taxonomy of Viruses (ICTV), several publications have used the terms '*Megaviridae*' or 'extended *Mimiviridae*' to refer to 'Mesomimivirinae' [40]. Here we will use the term *Mimiviridae* to encompass the formally recognized *Mimiviridae* as well as the proposed 'extended *Mimiviridae*' subfamily that are often considered phycodnaviruses. Proposed members of the 'Mesomimivirinae' include Organic Lake phycodnaviruses (OLPVs), *Aureococcus anophagefferens* virus (AaV), *Chrysochromulina ericina* virus (CeV), *Phaeocystis pouchetii* virus (PpV), *Pyramimonas orientalis* virus (PoV), and Group I *Phaeocystis globosa* viruses (PgVs) [39,41-44]. Based on these recent developments, we manually curated the affiliation of viruses that belong to the proposed Mesomimivirinae group but were annotated as

Phycodnaviridae in the current NCBI scheme (accessed April 2018), as viruses within the *Mimiviridae* (Table S1).

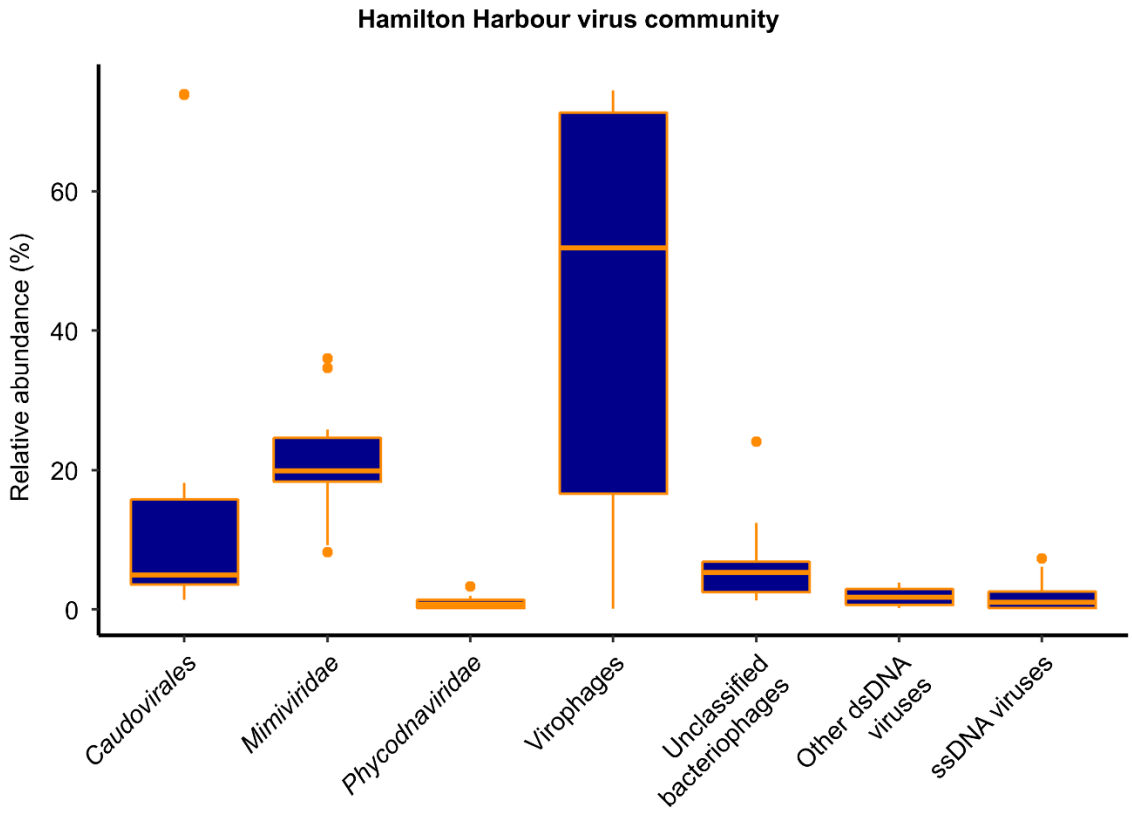
A boxplot of the overall virus community was generated using the "ggplot2" package [45] in RStudio. Bray-Curtis dissimilarity with unweighted pair group method with arithmetic mean (UPGMA) clustering was used to determine which samples were most similar based on relative abundances of different virus groups. The relationship between the relative abundance of virus groups, environmental parameters (dissolved oxygen, pH, chlorophyll a, temperature, Secchi depth, and redox potential), and sites, was statistically tested using canonical correspondence analysis (CCA); tests of 10,000 permutations were used to compute the significance of the model and the variables. Cluster analysis and CCA were computed using the "vegan" package [46] in RStudio. For samples collected on July 30th, and September 10th and 24th, data were available for other factors including: ammonia, chloride, fluoride, sulfate, particulate organic carbon, particulate organic nitrogen, nitrate and nitrite, total dissolved nitrogen, total phosphorus, total dissolved phosphorus, total particulate phosphorus, and soluble reactive phosphorus. Separate CCA models were tested for the entire dataset and the subset of dates for which additional data were available.

RESULTS

Virus community composition in Hamilton Harbour

Diverse virus contigs were identified in Hamilton Harbour and some were classified within the virus families *Caudovirales, Mimiviridae, Phycodnaviridae*, virophages (*Lavidaviridae*), while others could only be classified as belonging to *unclassified bacteriophages*, *other dsDNA*

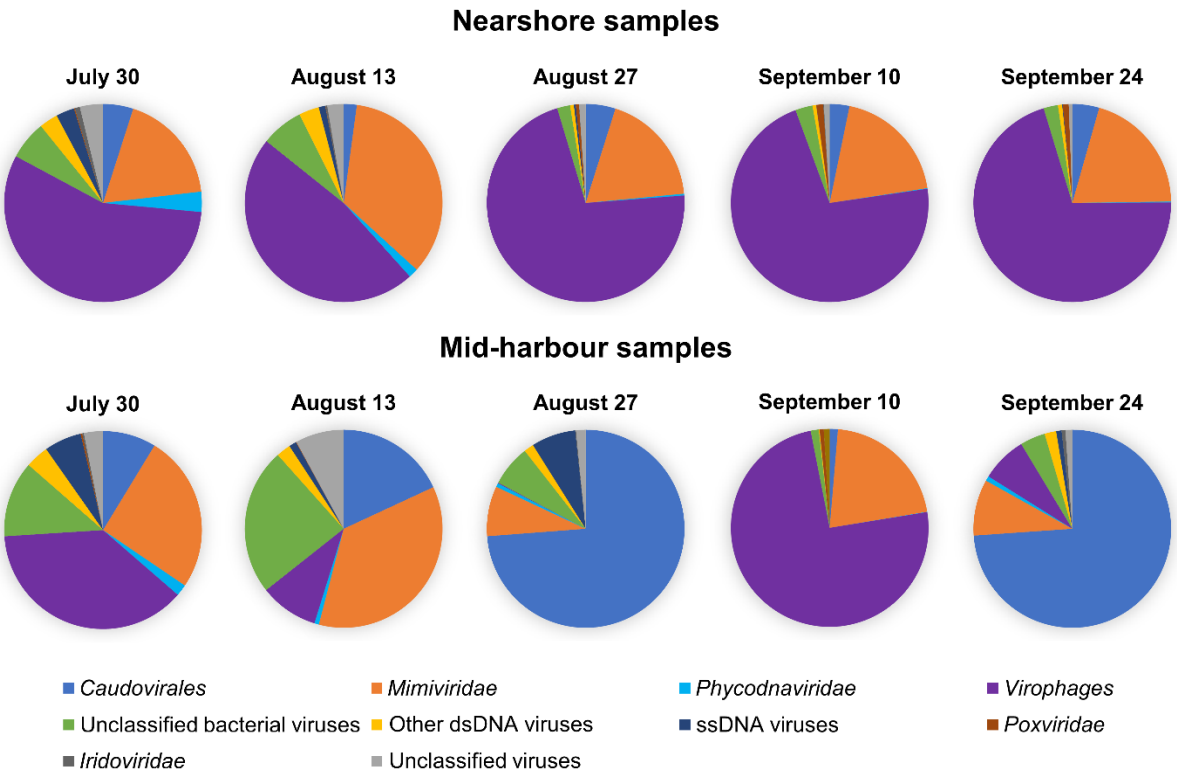
263 *viruses*, and *ssDNA viruses*. Henceforth, for the sake of brevity, we will name only virus groups
264 when referring to contigs annotated as viruses from those groups. Overall, virophages were the
265 most abundant (44.7% average abundance across all samples) but were also the most variable
266 (from 0.1% to 74.5% abundance). The *Mimiviridae* were the second most abundant group,
267 comprising an average of 21.1% of the virus community across all samples. Interestingly, though
268 they are intimately associated, the abundances of *Mimiviridae* (ranging from 8.2% to 36.0%
269 abundance) did not fluctuate to the same extent as the virophages (Figure 1). At the nearshore
270 site, virophages represented a large percentage of the virus community in every sample with
271 abundances between 47.4% and 71.7%, yet their abundances fluctuated widely at the mid-
272 harbour site ranging from 0.12% to 74.5%. *Mimiviridae* were consistently detected as a
273 substantial proportion of the community in the nearshore samples with relative abundances
274 ranging from 18.2% to 34.6% overall. Again, like the virophages, *Mimiviridae* abundances were
275 more variable at the mid-harbour site, comprising between 8.2% and 36.0% of all virus contigs.
276 Similarly, *Caudovirales* consistently represented less than 5.0% of the virus community in the
277 nearshore samples, but ranged from 1.3% to 73.9% at the mid-harbour site. *Phycodnaviridae*
278 were a minor component of the virus community in all samples, representing only 0.09% to 3.3%
279 at the nearshore site and 0.05% to 1.9% at the mid-harbour site (Figure 2).



280

281 **Figure 1:** Boxplot of the overall virus community in Hamilton Harbour based on all 10

282 metagenomes from the 5 sampling dates at the nearshore and mid-harbour sites.

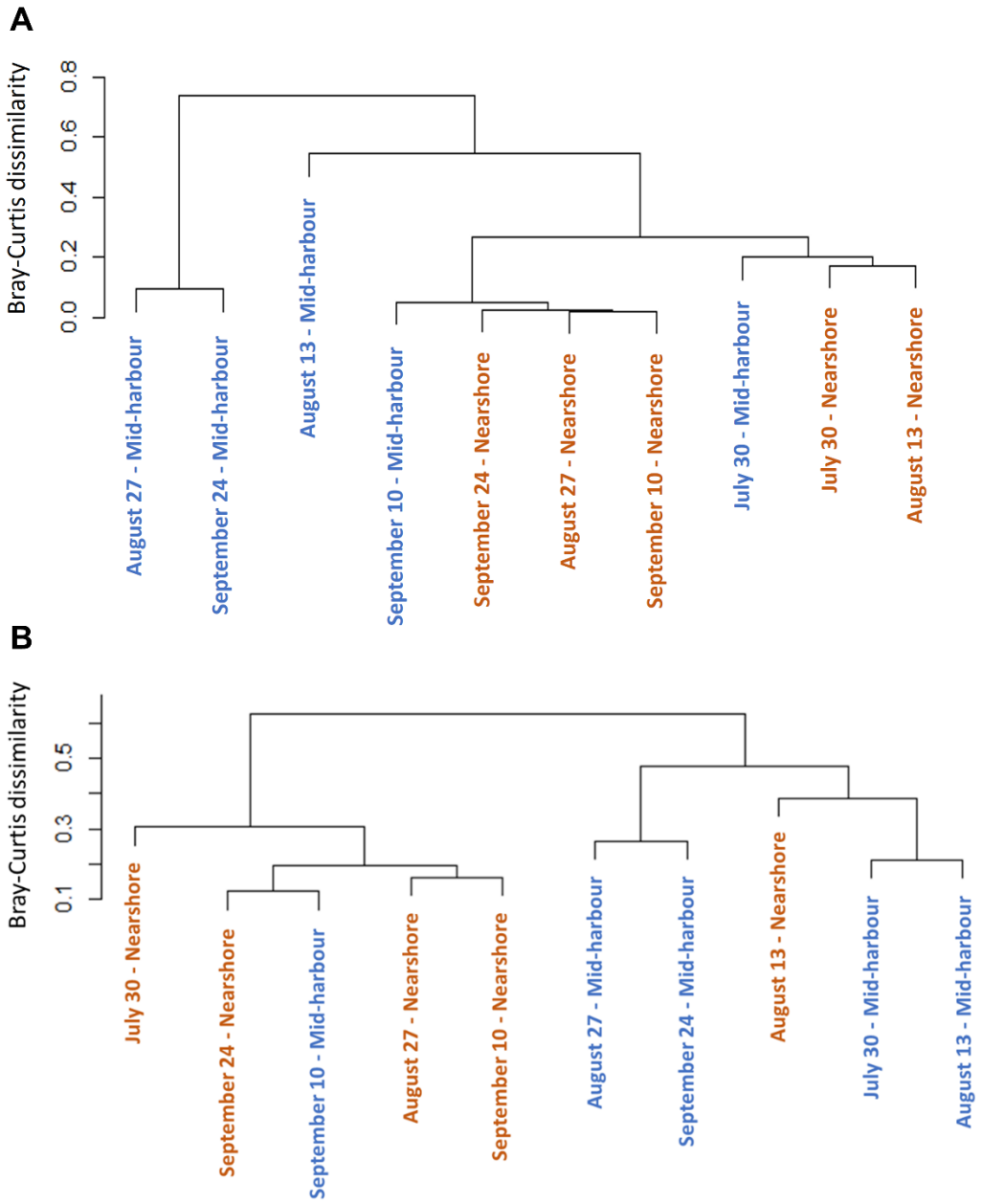


283

284 **Figure 2:** Individual virus communities for each of the 10 metagenomes.

285

286 In general, the abundance of different groups of viruses was more variable in the mid-harbour
287 compared to the nearshore samples. Sample similarity based on community composition was
288 assessed using UPGMA clustering of a Bray-Curtis dissimilarity matrix and reinforced this
289 contrast of the nearshore and mid-harbour sites (Figure 3a). Though the nearshore and mid-
290 harbour samples did not form distinct clusters overall, the nearshore samples clustered more
291 closely together than the mid-harbour samples. The community composition at the two sites
292 clustered together on July 30th and September 10th, but resolved to distinct clusters on August
293 13th, August 27th and September 24th.



294

295 **Figure 3:** Dendrogram of cluster analysis based on Bray-Curtis dissimilarity index of the **a)**

296 overall virus community and **b)** *Mimiviridae* community. Nearshore samples are coloured orange

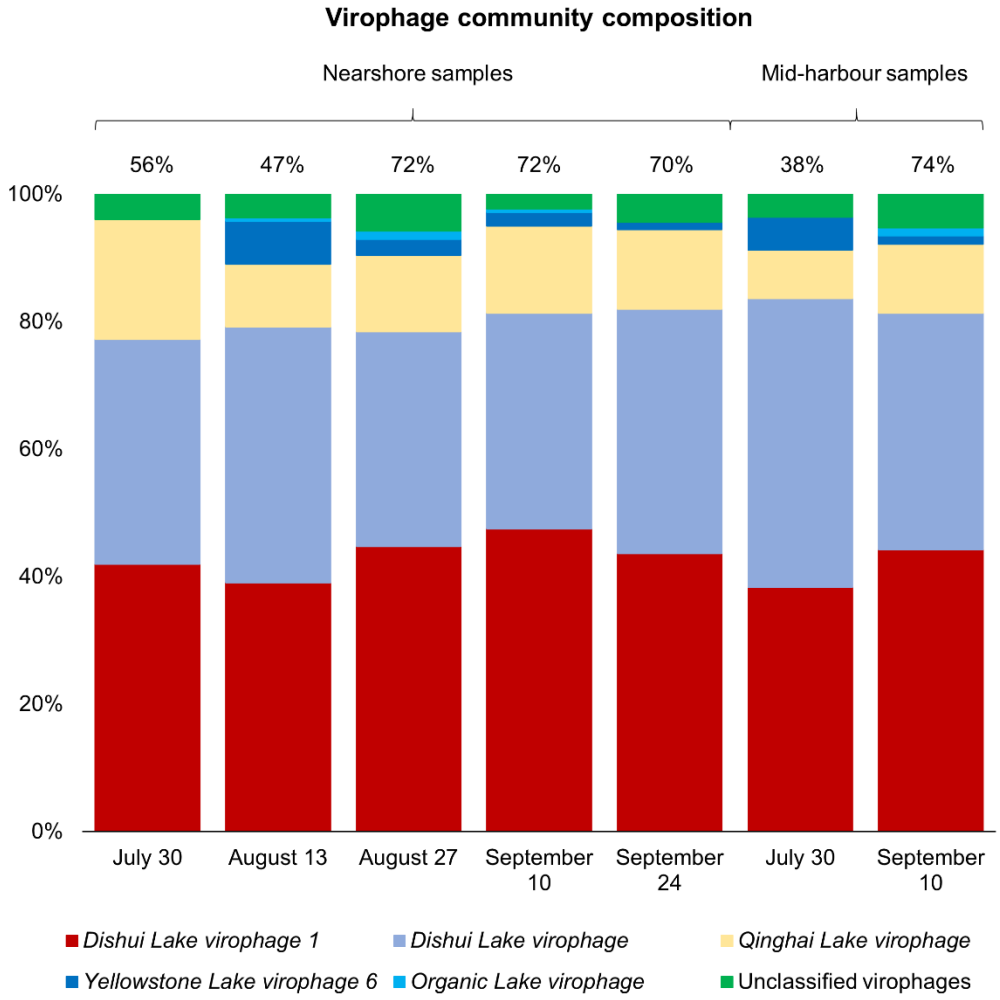
297 and mid-harbour samples are coloured blue.

298

299

Mimiviridae community composition in Hamilton Harbour

Diverse *Mimiviridae* contigs were detected in Hamilton Harbour, including representatives of all subgroups and proposed subgroups. Most notably, the proposed 'Klosneuvirinae' subfamily [47] comprised a large proportion of the *Mimiviridae* community, ranging from 17.5% to 79.0% across all samples, and representing an average of 67.4% and 41.7% of the *Mimiviridae* community at the nearshore site and mid-harbour site, respectively. In general, the nearshore and mid-harbour sites appeared to host distinct *Mimiviridae* communities, a notion supported the Bray-Curtis dissimilarity clustering analysis (Figure 3b); samples from the two sites clustered separately, with the exceptions of September 10th at the mid-harbour site and August 13th at the nearshore site. In all nearshore samples and on September 10th at the mid-harbour site, *Indivirus ILV1* were the most abundant representatives of the *Mimiviridae* community, while *Chrysochromulina ericinia* viruses (CeV*)* were the most abundant *Mimiviridae* in all mid-harbour samples except on September 10th (Figure 4) when *Indivirus ILV1* were again dominant.



314

315 **Figure 4:** Virophage community composition in samples where virophages represent >10% of
316 the total virus community. Numbers above bars are percentages of the total virus community that
317 the entire bar represents.

318

319 Of the proposed subfamily 'Mesomimivirinae', Organic Lake phycodnavirus, Yellowstone Lake
320 mimivirus and *Chrysochromulina ericinia* virus were the most abundant. The "Other
321 *Mimiviridae*" category was created for *Mimiviridae* that were present in only a few of the
322 samples and always at less than 2.0% of the community, and included the following
323 Mesomimivirinae members: *Aureococcus anophagefferens*, *Phaeocystis pouchetii*, *Pyramimonas*

orientalis, and *Phaeocystis globosa*. Also placed in the "Other *Mimiviridae*" category were *Acanthamoeba polyphaga* mimivirus, unclassified *Megaviridae*, Megavirus Iba, Powai Lake megavirus, and Mimivirus AB-566-O17, which were detected in 6 of the samples at less than 6.0% relative abundance.

Virophage community composition in Hamilton Harbour

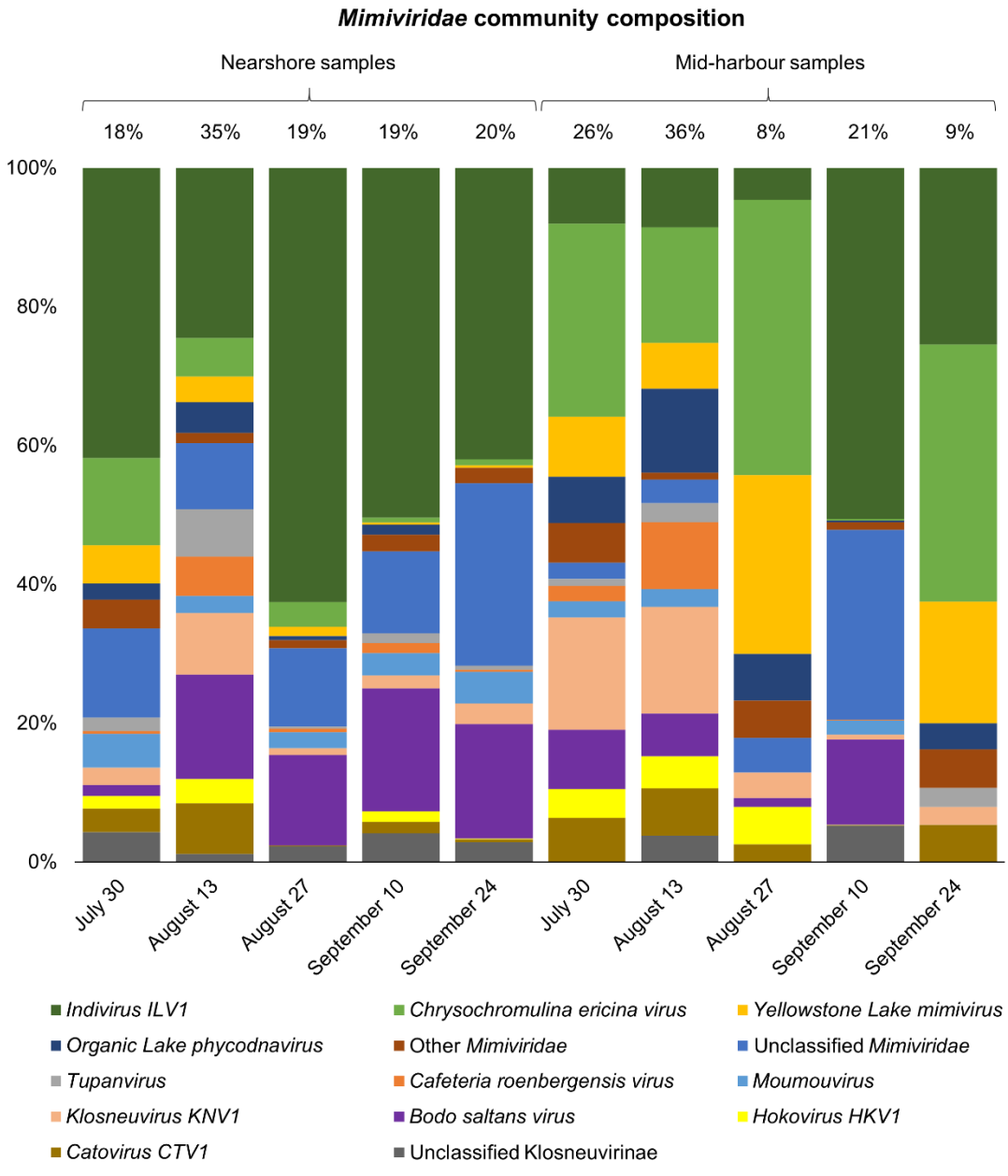
Given the similarity of virus community composition in the nearshore samples, we explored whether this consistency was upheld at the level of discrete taxa in the most abundant group, the virophages. Across all nearshore samples, virophage community composition was very similar regardless of date and contigs were annotated as five types of virophages: Dishui Lake virophage, Dishui Lake virophage 1, Yellowstone Lake virophage 6, Qinghai Lake virophage, and Organic Lake virophage (Figure 5). Across all samples, the majority of virophage contigs were most similar to Dishui Lake virophage 1 (43.4% average) and Dishui Lake virophage (36.3% average). Qinghai Lake virophages comprised 13.3% of the communities on average, and the Yellowstone Lake virophages were observed at lower abundances averaging only 2.5% of the virophage community. Least abundant were the Organic Lake virophages, which were detected at 0.5% abundance on average and were only detectable in 3 of the 5 nearshore samples. Unclassified virophage contigs were also detected at low abundances on all dates, averaging 4.0% of the virophage community. Overall, from July 30th to September 24th the virophage community composition at the nearshore site was remarkably similar. In contrast to virophage communities at nearshore sites, the mid-harbour samples from July 30th and September 10th were the only samples with virophage populations comprising >10% of the total virus community. Again, both samples were dominated by the Dishui Lake virophages (Dishui Lake virophage 1

347 and Dishui Lake virophage), which comprised about 80% of the total virophage population.

348 Regardless of site or date, the contigs most closely resembled virophages originally identified in

349 Dishui Lake, China (Figure 5).

350



351

352 **Figure 5:** *Mimiviridae* community composition at the nearshore and mid-harbour sites from July
353 30th to September 24th. Numbers above bars are percentages of the total virus community that the
354 entire bar represents.

355

356 *Influence of environmental parameters on virus community composition*

357 A canonical correspondence analysis (CCA) was performed to assess the influence of pH,
358 temperature, redox potential, dissolved oxygen, Secchi depth and chlorophyll a on the viral
359 groups at each site. The CCA model explained 47.8% ($F = 7.12$, $Pr(>F) = 0.005$) of the
360 variability in the data. A test of 10,000 permutations revealed that chlorophyll a concentration
361 was the only significant environmental parameter of those assessed. Temperature, pH, dissolved
362 oxygen, Secchi depth, and redox potential were not significant explanatory factors of changes in
363 virus community composition and relative abundances between samples. However, there was a
364 strong inverse relationship of *Caudovirales* and chlorophyll a concentration. The August 27th and
365 September 24th mid-harbour samples, which are dominated by *Caudovirales* (>70%), were
366 similar in community composition (Figure 3a) and were negatively correlated with chlorophyll a.
367 All nearshore samples clustered closely together and were positively correlated with chlorophyll
368 a. Several parameters including ammonia, chloride, fluoride, sulfate, particulate organic carbon,
369 particulate organic nitrogen, nitrate and nitrite, total dissolved nitrogen, total phosphorus, total
370 dissolved phosphorus, total particulate phosphorus, and soluble reactive phosphorus were
371 measured only on July 30th, September 10th and September 24th at both sites; however, none were
372 significant explanatory variables of the virus communities on these dates. Interestingly, when the
373 influence of environmental variables on the *Mimiviridae* community was assessed, chlorophyll a
374 remained an explanatory variable, accounting for 37.8% of the variation in the *Mimiviridae*

375 community (F = 4.85, Pr(>F) = 0.018). Of the parameters measured only on July 30th, September
376 10th and September 24th, only particulate organic carbon (POC) significantly explained
377 differences in the *Mimiviridae* communities between samples. In a separate CCA model, a test of
378 719 permutations revealed that POC accounted for 59.6% of the variation in the *Mimiviridae*
379 community on July 30th, September 10th and September 24th (F = 5.90, Pr(>F) = 0.043).
380
381 DISCUSSION
382 *Influence of databases*
383 In this study we observed diverse virus communities that were both spatially and seasonally
384 variable, and often dominated by virophages. However, as usual for environmental
385 metagenomes, a large portion of sequences remained unclassified and were discarded at the
386 annotation step of the data processing pipeline. This is especially the case for viruses, which
387 have much fewer sequenced representatives in databases than prokaryotes or eukaryotes and are
388 more likely to remain unannotated. For example, in May 2018, RefSeq release 88 contained
389 >7500 viruses with >325,300 associated accessions (9648 nucleotide sequences and 315,742
390 protein sequences). In contrast, there were >50,400 bacteria with >100,583,400 associated
391 accessions (12,367,951 nucleotide sequences and 88,193,695 protein sequences) (data from:
392 ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-statistics/). An estimated <1% of the Earth's
393 virome has been discovered [48], and this underrepresentation in databases has a large influence
394 on the reported diversity of viruses in metagenomes. The NCBI-nr database referenced in this
395 study combines data from RefSeq as well as SwissProt, Protein Information Resource (PIR),
396 Protein Databank (PDB), Protein Research Foundation (PRF), and GenBank. A combination of
397 curated and non-curated sequences, the NCBI-nr database includes data from the assembly of

398 environmental metagenomes including those deposited by private institutions. The use of this
399 database as opposed to a curated database drastically increases the size of the reference database,
400 increasing the likelihood of annotation, which is especially valuable for identifying taxa with few
401 sequenced representatives.
402
403 Within viral sequence databases, certain groups of viruses have many sequenced representatives,
404 while others are represented by only a few sequences. For example, there are >2000 complete
405 *Caudovirales* genomes, but only 7 complete virophage (*Lavidaviridae*) genomes (data from
406 https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi; retrieved December 2018). This
407 bias results in a higher chance of detecting *Caudovirales* than virophages, which are more likely
408 to be unannotated and discarded in downstream analyses. Additionally, viral groups with more
409 sequenced representation are more likely to have accurate annotations at the level of discrete
410 taxa. Hence, the disproportionately high abundance of virophages detected in Hamilton Harbour
411 samples compared to reference databases reinforces the discovery of their relatively high
412 representation in this system, and likely other freshwater environments.
413
414 *Limitations of inferring virus presence and abundance from metagenomic datasets*
415 This research infers the presence and abundance of viruses based on the relative abundances of
416 assembled contigs from shotgun sequencing. It is important to note that the presence of contigs
417 annotated as viruses may not necessarily indicate their presence. The intimate evolutionary
418 relationship between viruses and their hosts has resulted in an abundance of viral genes present
419 in host genomes. Unless viral homologues in host genomes are specifically annotated as such,
420 genetic similarity searches may incorrectly annotate them as viruses [49]. However, upon

421 examining the top 10 hits for individual contigs, there were few instances that contained a
422 combination of bacteria and viruses; in most cases the top hits were all within the same virus
423 family.
424
425 The concept of the "virocell" was introduced to distinguish between the living and non-living
426 segments of the classical virus lifestyle [50]. The non-living portion of the virus lifecycle is the
427 inactive virion that may become activated only if it contacts and infects a permissive host cell,
428 while the living manifestation of a virus is the infectious and metabolically active intracellular
429 portion of the virus lifecycle (i.e. the virocell). Counting viral contigs without distinguishing
430 between active and inactive states may lead to an overestimation of the influence of viruses in a
431 given environment if the data are not considered in light of this complex relationship between
432 viruses and their hosts. Metagenomic research is limited by the inability to distinguish between
433 viruses that are actively infecting hosts and virions that are inactive and simply adsorbed to cells
434 or particles. Furthermore, in this study we could not distinguish between viruses integrated into
435 host genomes and those present as virions. However, since DNA was extracted from the >0.22
436 um fraction rather than the traditional <0.22 um fraction, it is less likely that the virus contigs
437 captured were derived from free virions in the water column. Most free virions, except some of
438 the largest NCLDVs, would have passed through the filters, and only those that were adsorbed to
439 cells or other particulate matter would have been captured during sample collection. Though
440 virophages are small and could pass through the filters if present as free particles in the water
441 column, it is also possible that several virophages could be adsorbed to individual giant virus
442 hosts. For example, the virophage Sputnik is frequently observed to be situated within the fibrils
443 of its *Mamavirus* host and is speculated to use these fibrils in order to gain entry into the viral

444 host [51]. More research is required to confirm virophage host associations and methods of entry
445 in order to adequately assess whether virophages commonly adsorb to their giant virus hosts, and
446 what influence this may have on the interpretation of metagenomic data gathered in studies such
447 as this. Regardless, the high abundance of virophages in the majority of Hamilton Harbour
448 samples is indicative of their potential influence on the *Mimiviridae* and eukaryotic host
449 communities in the harbour, and likely many other freshwater environments as well.
450
451 Though these observations contrast metagenomic studies in other freshwater lakes, it remains
452 challenging to compare virus communities between studies due to differences in sampling,
453 sequencing and data processing. Even the most widely-used analysis tools for metagenomic data
454 yield different results [52], highlighting the need for standardized analysis tools and pipelines in
455 order to facilitate ecologically relevant comparisons between studies.
456
457 *Physiochemical factors*
458 Environmental variables that typically explain variation in virus communities on local scales,
459 such as dissolved oxygen and temperature [6], were not significant explanatory variables of our
460 data. This may be a reflection of virus dependence on host populations and the wide host ranges
461 of virus families. For example, the *Mimiviridae* infect hosts ranging from heterotrophic protists
462 like amoeba to photosynthetic protists, or algae, and because key resources for these eukaryotic
463 microorganisms differ (DOC versus light), it might be anticipated that they respond differently to
464 certain environmental factors. Nevertheless, the CCA model with chlorophyll a as a constrained
465 variable explained 47.8% of the variation in the data, while other environmental parameters were
466 insignificant. Therefore, it appears that fluctuations in algal biomass (inferred from chlorophyll a

concentration) play a major role in shaping the virus communities in Hamilton Harbour. For example, chlorophyll a was inversely related to *Caudovirales* relative abundance, indicating that most *Caudovirales* contigs were derived primarily from phages of heterotrophic bacteria. The mid-harbour samples from August 27th and September 24th had undetectable or very low (0.1 µg/L) chlorophyll a concentrations, respectively, and the virus communities in these samples were dominated by *Caudovirales* and clustered closely together in both the CCA and the Bray-Curtis dissimilarity analyses. In contrast, the lowest chlorophyll a concentration noted for a nearshore sample was 2.7 µg/L, and the virus community in this sample (August 13th) was dominated by virophages and *Mimiviridae*. It is notable that the nearshore site is closer in proximity to wastewater effluents from Spencer Creek than the mid-harbour site. Although improvements have been made to the wastewater treatment plants surrounding Hamilton Harbour, effluents still contain phosphorus levels above targets set in the Hamilton Harbour Remedial Action Plan (HHRAP) [53]. These nutrient and organic carbon inputs might stimulate microbial growth at the nearshore site, in turn influencing virus communities.

The CCA of different *Mimiviridae* taxa against environmental parameters revealed that chlorophyll a remained a significant explanatory variable of the *Mimiviridae* community, reflecting their wide host range which includes both photosynthetic and non-photosynthetic hosts. For example, the algal virus CeV made up a large portion of the *Mimiviridae* community at the mid-harbour site, while *Indivirus ILV1*, a member of the proposed subfamily 'Klosneuvirinae' that were themselves assembled from metagenome sequences, was abundant in nearshore samples and is a suspected protist virus [47]. POC explained almost 60% of the variation in the *Mimiviridae* communities on July 30th, September 10th and September 24th. Since

490 POC data were only available for 6 of the 10 samples, it is unclear whether POC would remain a
491 significant explanatory variable if the model included data from all samples.
492
493 Hamilton Harbour is known to be a highly variable system [54] that experiences regular seiches
494 [55] and exchange flows with Lake Ontario, especially during the summer [27]. Circulation and
495 mixing in the harbour are primarily controlled by prevailing winds [56]. Modelling of summer
496 circulation patterns in Hamilton Harbour demonstrated the occurrence of a large eddy in the
497 middle of the harbour at the location of our mid-harbour site, while the nearshore site was
498 situated on the perimeter of a smaller eddy located at western end of the bay near Cootes
499 Paradise [57]. The location of the sampling sites with respect to these eddies and Hamilton
500 Harbour circulation may explain differences in virus community composition observed at the
501 different sites on the same day. Hamilton Harbour is a major Canadian shipping port and large
502 ships entering the harbour from Lake Ontario likely pass through the mid-harbour site on route to
503 the port on the southern shore. Shipping traffic is one of the many factors potentially affecting
504 virus community variability on the sampling dates that could not be considered in the present
505 study. Given the high variability of Hamilton Harbour, it is perhaps unsurprising that the virus
506 communities varied widely between sampling sites and dates. More unexpected was the
507 relatively stable community throughout the mid-summer to late fall at the nearshore site, when
508 bacterial and phytoplankton populations have been observed to be highly dynamic [58-60].
509
510 *The virophages*
511 Virophages are small dsDNA viruses that co-infect eukaryotic hosts with giant dsDNA viruses
512 [61]. Their co-infection has been shown to reduce giant virus fitness, thereby increasing survival

513 of the cellular host [62,63]. Virophages were discovered only a decade ago [63], and little is
514 known about their ecology. There is evidence to suggest that virophage-induced reduction of
515 algal host mortality leads to longer and more frequent algal blooms [64], highlighting the
516 ecological importance of virophages and their potential relevance to our eutrophic study site.
517
518 Virophage abundance in freshwater eutrophic lakes varies widely and has been reported as
519 highly abundant in some environments [26], yet only detected at low abundances (<0.05%) [20],
520 or even undetectable in other lakes [22]. Virophage abundances and distributions in lakes do not
521 have clear seasonal or annual patterns of abundance, nor are there established relationships
522 between dominant types of virophages in different environments [26]. In contrast to previous
523 studies of viruses in eutrophic freshwater lakes, virophages were dominant in most samples from
524 Hamilton Harbour. About 80% of the virophage community was annotated as Dishui Lake
525 virophages in 7 of 10 samples regardless of date or site. Dishui Lake, China is a eutrophic
526 freshwater lake that is more similar to Hamilton Harbour than the lakes where other types of
527 virophages were observed such as Organic Lake, an Antarctic hypersaline lake, Qinghai Lake, a
528 saline endorheic basin, or Yellowstone Lake, a lake that receives geothermal inputs. Due in part
529 to their relatively recent discovery, sequence databases are limited with respect to the amount of
530 information available for virophages, hence, specific virophage annotations in our samples may
531 not be accurate. It is likely that the large portion of the virophage community annotated as Dishui
532 Lake virophages are in fact a diverse array of virophages. The breakdown into discrete taxa did,
533 however, reveal which virophages in reference databases were most similar to those detected in
534 our dataset.
535

Other considerations that may impact the accuracy of virophage annotations include the presence of polintons/polintoviruses, polinton-like viruses and transpovirons (reviewed in [65]). Related to virophages, polintons/polintoviruses are self-synthesizing transposons that have genes homologous to virophages and giant viruses and are frequently integrated into eukaryotic genomes. Polinton-like viruses (PLV) are related to polintons and are commonly integrated into the genomes of green algae [65]. Mimiviruses can also be infected by transpovirons, which are small, dsDNA parasites that share homologous genes with virophages. During mimivirus reproduction, numerous transpovirons can accumulate within the host cytoplasm, within mimivirus particles, and within virophage particles. They also have the ability to integrate into the genomes of mimiviruses [66]. Considering their evolutionary relatedness and the presence of homologous genes in conjunction with the limitations of available databases, it would not be surprising to find that some of the virophages detected in our samples may in fact be polintons/polintoviruses, PLVs, or transpovirons. Further exploration and sequencing of virophage, polinton, PLV and transpoviron diversity is required to expand databases and improve accuracy of identification of these entities in metagenomic datasets. As a final comment on virophages, recent genome sequencing of a giant virus isolated from Lake Ontario revealed the presence of three putative virophages [67]. This giant virus and associated virophages infect the freshwater haptophyte *Chrysochromulina parva*; the *C. parva* virus is a close relative to the Hamilton Harbour mimivirus contigs annotated as *Chrysochromulina ericinia* virus. This observation supports the notion that the contigs assembled from Hamilton Harbour represent bona fide mimivirus-parasitizing virophages.

559 *Ecological relevance*

560 The importance of viruses in aquatic environments is well documented. They are estimated to
561 kill approximately 10% of the phytoplankton population and up to 50% of the bacterial
562 population in surface marine waters, with greater impacts in high nutrient environments [68].
563 Especially in eutrophic aquatic systems, viruses are more active and are hypothesized to control
564 host abundance, respiration and production [69]. They drive host community succession by
565 targeting and lysing abundant members of the community, allowing less competitive species to
566 thrive [70], and influencing fluctuations in dissolved and particulate organic matter pools.
567 Viruses also act as key gene transfer agents that permit host adaptation and drive the evolution of
568 microbial communities [71].

569

570 While marine viruses have been studied extensively over the past decade, freshwater viruses
571 have received relatively little attention. Fundamental aspects of freshwater virus ecology, such as
572 their distribution and patchiness in freshwater environments remain unknown. To our
573 knowledge, this research is the first report of virus communities in Hamilton Harbour, and one of
574 the few studies of virus communities in the Great Lakes. On some dates (e.g. September 24th)
575 virus community composition was very different over the 3-km distance between the sites, while
576 on other dates (e.g. September 10th), the community composition was very similar. At the level
577 of virus families, relative abundances fluctuated at the mid-harbour site much more than the
578 nearshore site, highlighting the high diversity of viruses on local scales and the impacts of small-
579 scale environmental differences.

580

581 Since the recent discoveries of the *Mimiviridae* and their virophages, few studies have looked at
582 the relative abundances of these groups over the duration of a season. We captured fluctuations
583 in the *Mimiviridae* community over the summer and observed vastly different communities on
584 the same date at different locations in the harbour. The stability of the virophage community
585 abundance throughout the nearshore samples despite the changing *Mimiviridae* community was
586 unexpected as *Mimiviridae* are the only known virus hosts of virophages [72]. While the
587 nearshore and mid-harbour samples generally contained distinct *Mimiviridae* communities, the
588 virophage community composition remained consistent. This highlights the complexity of these
589 ecological relationships, and the gaps in our understanding of how these intimately associated
590 viruses interact. The stability of virophage community composition despite the fluctuating
591 Mimiviridae community appears to support the hypothesis that virophages and their *Mimiviridae*
592 hosts are not connected solely by infection. The notion that virophages may enter eukaryotic host
593 cells independently, remaining latent until infection by a giant virus, where they then compete
594 with the giant virus for its replication machinery [72] is a possible explanation of the wide range
595 of abundances in the virophage community while the *Mimiviridae* community remained
596 relatively stable.
597
598 Given that Hamilton Harbour is known to support high algal biomass in the summer and early
599 autumn months, it was surprising that the *Phycodnaviridae* were observed at low relative
600 abundances in all samples collected for this study. Instead, *Mimiviridae* appeared to be the
601 dominant algal viruses in Hamilton Harbour. The most common algae-infecting *Mimiviridae*
602 were CeV, which were especially abundant at the mid-harbour site. The “Other dsDNA viruses”
603 category did not contribute more than 4% in any individual sample and included mostly

604 *Mimiviridae*. The small percentage of ssDNA viruses that were detected represent only those that
605 were replicating in cells in the dsDNA form, since only dsDNA was targeted in the library
606 preparation and sequencing. Therefore, the ssDNA viruses may be underrepresented compared to
607 other virus groups which were detected as inactive virions adsorbed to particles in addition to
608 active viruses replicating within cells.
609
610 CONCLUSIONS
611 Overall, Hamilton Harbour metagenomes included a diverse array of viruses ranging from large
612 dsDNA *Mimiviridae* to small ssDNA viruses. Relative abundances of virus families varied
613 widely over relatively small spatial scales within the harbour, with higher consistency in the
614 nearshore samples compared to the mid-harbour samples. Virophage relative abundances ranged
615 widely across all samples and were the most abundant virus family in most samples. Though
616 *Mimiviridae* are presumably intimately associated with virophages, their abundances did not
617 fluctuate to the same extent as the virophages. A wide diversity of *Mimiviridae* were detected in
618 the samples, and the two sites appeared to host distinct *Mimiviridae* communities. The
619 abundances of discrete virophage taxa were remarkably stable despite the dissimilar *Mimiviridae*
620 communities at the two sites, highlighting our limited understanding of how these intimately
621 associated viruses interact. Equally unexpected was the low abundance of *Caudovirales* in most
622 samples, contrasting other studies of freshwater virus communities. *Phycodnaviridae* abundances
623 were also surprisingly low in all samples despite Hamilton Harbour's capacity to support high
624 algal biomass during the summer and autumn months, suggesting that *Mimiviridae* are the
625 dominant algae-infecting viruses in this system. These findings provide insight into virus
626 community structures in freshwater environments, expanding the documented diversity of

627 freshwater virus communities, highlighting the potential ecological importance of virophages,
628 and revealing distinct communities over small spatial scales.
629

630 SUPPLEMENTARY MATERIALS

631 **Table S.1:** List of original taxonomic annotations and re-assigned taxonomic affiliations of virus
632 contigs in Hamilton Harbour metagenomes.
633

634 ACKNOWLEDGMENTS

635 This work was supported in part by an NSERC Discovery Grant (#RGPIN-2016-06022) awarded
636 to S.M.S. and an Ontario Graduate Scholarship (OGS) awarded to C.N.P. Hamilton Harbour
637 sampling, DNA extraction and metagenomic sequencing was funded by Environment Canada.
638 Environmental parameters were collected and measured by researchers at York University (Dr.
639 Lewis Molot's group) and Environment Canada (Dr. Susan Watson's group). These samples
640 were processed by the staff at the National Laboratory for Environmental Testing (NLET). Our
641 appreciation to Dr. Daniel Huson for correspondence regarding data processing in DIAMOND
642 and MEGAN6-LR.
643

644 AUTHOR CONTRIBUTIONS

645 Conceptualization, C.N.P. and S.M.S.; Methodology, C.N.P., R.R.F., R.S., and S.M.S;
646 Validation, C.N.P. and S.M.S.; Formal Analysis, C.N.P. and S.M.S; Investigation, C.N.P.,
647 R.R.F., R.S., and S.M.S.; Resources, R.R.F. and S.M.S.; Data Curation, C.N.P. and S.M.S.;

Writing – Original Draft Preparation, C.N.P. and S.M.S; Writing – Review & Editing, C.N.P., R.R.F., R.S., and S.M.S.; Visualization, C.N.P. and S.M.S.; Supervision, R.R.F. and S.M.S.; Project Administration, R.R.F. and S.M.S.; Funding Acquisition, R.R.F. and S.M.S.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. Jacquet, S.; Miki, T.; Noble, R.; Peduzzi, P.; Wilhelm, S. Viruses in aquatic ecosystems: important advancements of the last 20 years and prospects for the future in the field of microbial oceanography and limnology. *Advances in Oceanography and Limnology* **2010**, *1*, 97-141, doi:10.1080/19475721003743843.
2. Bergh, O.; Borsheim, K.Y.; Bratbak, G.; Heldal, M. High abundance of viruses found in aquatic environments. *Nature* **1989**, *340*, 467-468, doi:10.1038/340467a0.
3. Wommack, K.E.; Bhavsar, J.; Polson, S.W.; Chen, J.; Dumas, M.; Srinivasiah, S.; Furman, M.; Jamindar, S.; Nasko, D.J. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences* **2012**, *6*, 421-433, doi:10.4056/sigs.2945050.
4. Breitbart, M.; Salamon, P.; Andresen, B.; Mahaffy, J.M.; Segall, A.M.; Mead, D.; Azam, F.; Rohwer, F. Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America* **2002**, *99*, 14250-14255, doi:10.1073/pnas.202488399.
5. Brum, J.R.; Sullivan, M.B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nature Reviews Microbiology* **2015**, *13*, 147-159, doi:10.1038/nrmicro3404.
6. Brum, J.R.; Ignacio-Espinoza, J.C.; Roux, S.; Doulcier, G.; Acinas, S.G.; Alberti, A.; Chaffron, S.; Cruaud, C.; de Vargas, C.; Gasol, J.M., et al. Patterns and ecological drivers of ocean viral communities. *Science* **2015**, *348*, doi:10.1126/science.1261498.
7. de Vargas, C.; Audic, S.; Henry, N.; Decelle, J.; Mahe, F.; Logares, R.; Lara, E.; Berney, C.; Le Bescot, N.; Probert, I., et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* **2015**, *348*, doi:10.1126/science.1261605.
8. Lima-Mendez, G.; Faust, K.; Henry, N.; Decelle, J.; Colin, S.; Carcillo, F.; Chaffron, S.; Ignacio-Espinosa, J.C.; Roux, S.; Vincent, F., et al. Determinants of community structure in the global plankton interactome. *Science* **2015**, *348*, doi:10.1126/science.1262073.

9. Sunagawa, S.; Coelho, L.P.; Chaffron, S.; Kultima, J.R.; Labadie, K.; Salazar, G.; Djahanschiri, B.; Zeller, G.; Mende, D.R.; Alberti, A., et al. Structure and function of the global ocean microbiome. *Science* **2015**, *348*, doi:10.1126/science.1261359.
10. Villar, E.; Farrant, G.K.; Follows, M.; Garczarek, L.; Speich, S.; Audic, S.; Bittner, L.; Blanke, B.; Brum, J.R.; Brunet, C., et al. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* **2015**, *348*, doi:10.1126/science.1261447.
11. Breitbart, M.; Rohwer, F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol.* **2005**, *13*, 278-284, doi:10.1016/j.tim.2005.04.003.
12. Hingamp, P.; Grimsley, N.; Acinas, S.G.; Clerissi, C.; Subirana, L.; Poulain, J.; Ferrera, I.; Sarmento, H.; Villar, E.; Lima-Mendez, G., et al. Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J* **2013**, *7*, 1678-1695, doi:10.1038/ismej.2013.59.
13. Duarte, C.M. Seafaring in the 21st Century: The Malaspina 2010 Circumnavigation Expedition. *Limnol Oceanogr Bull* **2015**, *24*, 11-14. doi:10.1002/lob.10008.
14. Roux, S.; Brum, J.R.; Dutilh, B.E.; Sunagawa, S.; Duhaime, M.B.; Loy, A.; Poulos, B.T.; Solonenko, N.; Lara, E.; Poulain, J., et al. Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. *Nature* **2016**, *537*, 689-693, doi:10.1038/nature19366.
15. Djikeng, A.; Kuzmickas, R.; Anderson, N.G.; Spiro, D.J. Metagenomic Analysis of RNA Viruses in a Fresh Water Lake. *Plos One* **2009**, *4*, doi:10.1371/journal.pone.0007264.
16. Hewson, I.; Barbosa, J.G.; Brown, J.M.; Donelan, R.P.; Eaglesham, J.B.; Eggleston, E.M.; LaBarre, B.A. Temporal Dynamics and Decay of Putatively Allochthonous and Autochthonous Viral Genotypes in Contrasting Freshwater Lakes. *Applied and Environmental Microbiology* **2012**, *78*, 6583-6591, doi:10.1128/aem.01705-12.
17. Roux, S.; Enault, F.; Robin, A.; Ravet, V.; Personnic, S.; Theil, S.; Colombet, J.; Sime-Ngando, T.; Debroas, D. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *Plos One* **2012**, *7*, doi:10.1371/journal.pone.0033641.
18. de Carcer, D.A.; Lopez-Bueno, A.; Pearce, D.A.; Alcami, A. Biodiversity and distribution of polar freshwater DNA viruses. *Science Advances* **2015**, *1*, doi:10.1126/sciadv.1400127.
19. Mohiuddin, M.; Schellhorn, H.E. Spatial and temporal dynamics of virus occurrence in two freshwater lakes captured through metagenomic analysis. *Frontiers in Microbiology* **2015**, *6*, doi:10.3389/fmicb.2015.00960.
20. Green, J.C.; Rahman, F.; Saxton, M.A.; Williamson, K.E. Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern Virginia, USA. *Aquatic Microbial Ecology* **2015**, *75*, 117-128, doi:10.3354/ame01752.
21. Sible, E.; Cooper, A.; Malki, K.; Bruder, K.; Watkins, S.C.; Fofanov, Y.; Putonti, C. Survey of viral populations within Lake Michigan nearshore waters at four Chicago area beaches. *Data Brief* **2015**, *5*, 9-12, doi:10.1016/j.dib.2015.08.001.
22. Skvortsov, T.; de Leeuwe, C.; Quinn, J.P.; McGrath, J.W.; Allen, C.C.; McElarney, Y.; Watson, C.; Arkhipova, K.; Lavigne, R.; Kulakov, L.A. Metagenomic Characterisation of the Viral Community of Lough Neagh, the Largest Freshwater Lake in Ireland. *PLoS One* **2016**, *11*, e0150361, doi:10.1371/journal.pone.0150361.
23. Ge, X.; Wu, Y.; Wang, M.; Wang, J.; Wu, L.; Yang, X.; Zhang, Y.; Shi, Z. Viral metagenomics analysis of planktonic viruses in East Lake, Wuhan, China. *Virol Sin* **2013**, *28*, 280-290, doi:10.1007/s12250-013-3365-y.

24. de Carcer, D.A.; Lopez-Bueno, A.; Alonso-Lobo, J.M.; Quesada, A.; Alcami, A. Metagenomic analysis of lacustrine viral diversity along a latitudinal transect of the Antarctic Peninsula. *Fems Microbiology Ecology* **2016**, *92*, doi:10.1093/femsec/fiw074.
25. Lopez-Bueno, A.; Tamames, J.; Velazquez, D.; Moya, A.; Quesada, A.; Alcami, A. High Diversity of the Viral Community from an Antarctic Lake. *Science* **2009**, *326*, 858-861, doi:10.1126/science.1179287.
26. Roux, S.; Chan, L.K.; Egan, R.; Malmstrom, R.R.; McMahon, K.D.; Sullivan, M.B. Ecogenomics of virophages and their giant virus hosts assessed through time series metagenomics. *Nature Communications* **2017**, *8*, doi:10.1038/s41467-017-01086-2.
27. Lawrence, G.; Pieters, R.; Zaremba, L.; Tedford, T.; Gu, L.; Greco, S.; Hamblin, P. Summer exchange between Hamilton Harbour and Lake Ontario. *Deep Sea Research Part II: Topical Studies in Oceanography* **2004**, *51*, 475-487, doi:10.1016/j.dsr2.2003.09.002.
28. CPCS. Hamilton's Working Waterfront: Port of Hamilton Economic Impact Study. **2016.** Available online: https://economy.hamiltonport.ca/wp-content/uploads/2016/11/PortOfHamilton_Economic_Impact_Report_Oct_2016.pdf.
29. IJC. Report on the ongoing remedial and preventive efforts by responsible governments and organizations relative to restoring Hamilton Harbour. **1999.** Available online: http://ijc.org/php/publications/html/hamhar/hamharsa.html.
30. Poulton, D.J. Trace Contaminant Status of Hamilton Harbour. *Journal of Great Lakes Research* **1987**, *13*, 193-201, doi:10.1016/s0380-1330(87)71642-6.
31. ECCC. Canadian Environmental Sustainability Indicators: Restoring the Great Lakes Areas of Concern. **2018.** Available online: www.canada.ca/en/environment-climate-change/services/environmentalindicators/restoring-great-lakes-areas-concern.html.
32. Andrews, S. FastQC: a quality control tool for high throughput sequence data. **2010.** Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
33. Joshi, N.A.; Fass, J.N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. **2011**. Available online: https://github.com/najoshi/sickle.
34. Peng, Y.; Leung, H.C.M.; Yiu, S.M.; Chin, F.Y.L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **2012**, *28*, 1420-1428, doi:10.1093/bioinformatics/bts174.
35. Buchfink, B., Xie, C., & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **2015**, *12*, 59-63, doi:10.1038/nmeth.3176.
36. Huson, D.H.; Albrecht, B.; Bagci, C.; Bessarab, I.; Gorska, A.; Jolic, D.; Williams, R.B.H. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct* **2018**, *13*, doi:10.1186/s13062-018-0208-7.
37. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **2012**, *9*, 357-U354, doi:10.1038/nmeth.1923.
38. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079, doi:10.1093/bioinformatics/btp352.
39. Gallot-Lavallee, L.; Blanc, G.; Claverie, J.-M. Comparative Genomics of Chrysochromulina Ericina Virus and Other Microalga-Infecting Large DNA Viruses

Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. *Journal of Virology* **2017**, *91*, doi:10.1128/jvi.00230-17.

40. Maruyama, F.; Ueki, S. Evolution and Phylogeny of Large DNA Viruses, Mimiviridae and Phycodnaviridae Including Newly Characterized Heterosigma akashiwo Virus. *Frontiers in Microbiology* **2016**, *7*, doi:10.3389/micb.2016.01942.
41. Claverie, J.-M.; Abergel, C. Mimiviridae: An Expanding Family of Highly Diverse Large dsDNA Viruses Infecting a Wide Phylogenetic Range of Aquatic Eukaryotes. *Viruses-Basel* **2018**, *10*, doi:10.3390/v10090506.
42. Moniruzzaman, M.; LeCleir, G.R.; Brown, C.M.; Gobler, C.J.; Bidle, K.D.; Wilson, W.H.; Wilhelm, S.W. Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host-virus coevolution. *Virology* **2014**, *466*, 60-70, doi:10.1016/j.virol.2014.06.031.
43. Yutin, N.; Colson, P.; Raoult, D.; Koonin, E.V. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. *Virology Journal* **2013**, *10*, doi:10.1186/1743-422x-10-106.
44. Short, S.M., Staniewski, M. A., Chaban, Y. V., Long, A. M., & Wang, D. The Diversity of Viruses Infecting Eukaryotic Algae. In *Viruses of Microorganisms: Diversity, Molecular Biology and Application*; Hymen, P., & Abedon, S., Eds.; Caister Academic Press: Norfolk, UK, 2018; p. 211-244.
45. Wickham, H. *ggplot2: elegant graphics for data analysis*; Springer: New York, USA, 2009.
46. Oksanen, J.; Blanchet, F.G.; Friendly, M.; Kindt, R.; Legendre, P.; McGlinn, D.; Minchin, P.R.; O'Hara, R.B.; Simpson, G.L.; Solymos, P., et al. vegan: Community Ecology Package. *R package version 2.4-5* **2017**. Available online: https://CRAN.R-project.org/package=vegan.
47. Schulz, F.; Yutin, N.; Ivanova, N.N.; Ortega, D.R.; Lee, T.K.; Vierheilig, J.; Daims, H.; Horn, M.; Wagner, M.; Jensen, G.J., et al. Giant viruses with an expanded complement of translation system components. *Science* **2017**, *356*, 82-85, doi:10.1126/science.aal4657.
48. Mokili, J.L.; Rohwer, F.; Dutilh, B.E. Metagenomics and future perspectives in virus discovery. *Current Opinion in Virology* **2012**, *2*, 63-77, doi:10.1016/j.coviro.2011.12.004.
49. Forterre, P. The virocell concept and environmental microbiology. *Isme J.* **2013**, *7*, 233-236, doi:10.1038/ismej.2012.110.
50. Forterre, P. Manipulation of cellular syntheses and the nature of viruses: The virocell concept. *Comptes Rendus Chimie* **2011**, *14*, 392-399, doi:10.1016/j.crci.2010.06.007.
51. Desnues, C.; Raoult, D. Inside the Lifestyle of the Virophage. *Intervirology* **2010**, *53*, 293-303, doi:10.1159/000312914.
52. Lindgreen, S.; Adair, K.L.; Gardner, P.P. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific Reports* **2016**, *6*, doi:10.1038/srep19233.
53. Posedowski, B. *Dundas Wastewater Treatment Plant and Cootes Paradise*; City of Hamilton Public Works Department: Hamilton, Canada, 2018. Available online: https://pub-hamilton.escribemeetings.com/filestream.ashx?DocumentId=150416.
54. Haffner, G.D.; Harris, G.P.; Jarai, M.K. Physical variability and phytoplankton communities 3. Vertical structure in phytoplankton populations. *Archiv Fur Hydrobiologie* **1980**, *89*, 363-381.

55. Wu, J.; Tsanis, I.K.; Chiocchio, F. Observed currents and water levels in Hamilton Harbour. *Journal of Great Lakes Research* **1996**, *22*, 224-240, doi:10.1016/s0380-1330(96)70951-6.
56. Yerubandi, R.R.; Marvin, C.H.; Zhao, J. Application of a numerical model for circulation, temperature and pollutant distribution in Hamilton Harbour. *Journal of Great Lakes Research* **2009**, *35*, 61-73, doi:10.1016/j.jglr.2008.09.004.
57. Yerubandi, R.R.; Boegman, L.; Bolkhari, H.; Hiriart-Baer, V. Physical processes affecting water quality in Hamilton Harbour. *Aquatic Ecosystem Health & Management* **2016**, *19*, 114-123, doi:10.1080/14634988.2016.1165035.
58. Munawar, M.; Fitzpatrick, M. Microbial - Planktonic foodweb dynamics of a eutrophic Area of Concern: Hamilton Harbour. *Aquatic Ecosystem Health & Management* **2017**, *20*, 214-229, doi:10.1080/14634988.2017.1305865.
59. Munawar, M.; Fitzpatrick, M.; Niblock, H.; Kling, H.; Rozon, R.; Lorimer, J. Phytoplankton ecology of a culturally eutrophic embayment: Hamilton Harbour, Lake Ontario. *Aquatic Ecosystem Health & Management* **2017**, *20*, 201-213, doi:10.1080/14634988.2017.1307678.
60. Saati, R. Characterization of the Cyanobacterial Harmful Algal Bloom Community in Hamilton Harbour, Lake Ontario. MSc, University of Toronto, Canada, November 2016.
61. Krupovic, M.; Koonin, E.V. Self-synthesizing transposons: unexpected key players in the evolution of viruses and defense systems. *Current Opinion in Microbiology* **2016**, *31*, 25-33, doi:10.1016/j.mib.2016.01.006.
62. Claverie, J.M.; Abergel, C. Mimivirus and its Virophage. In *Annual Review of Genetics*, 2009; Vol. 43, pp. 49-66. doi: 10.1146/annurev-genet-102108-134255.
63. La Scola, B.; Desnues, C.; Pagnier, I.; Robert, C.; Barrassi, L.; Fournous, G.; Merchat, M.; Suzan-Monti, M.; Forterre, P.; Koonin, E., et al. The virophage as a unique parasite of the giant mimivirus. *Nature* **2008**, *455*, 100-U165, doi:10.1038/nature07218.
64. Yau, S.; Lauro, F.M.; DeMaere, M.Z.; Brown, M.V.; Thomas, T.; Raftery, M.J.; Andrews-Pfannkoch, C.; Lewis, M.; Hoffman, J.M.; Gibson, J.A., et al. Virophage control of antarctic algal host-virus dynamics. *Proc Natl Acad Sci U S A* **2011**, *108*, 6163-6168, doi:10.1073/pnas.1018221108.
65. Koonin, E.V.; Krupovic, M. Polintons, virophages and transpovirons: a tangled web linking viruses, transposons and immunity. *Current Opinion in Virology* **2017**, *25*, 7-15, doi:10.1016/j.coviro.2017.06.008.
66. Desnues, C.; La Scola, B.; Yutin, N.; Fournous, G.; Robert, C.; Azza, S.; Jardot, P.; Monteil, S.; Campocasso, A.; Koonin, E.V., et al. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci U S A* **2012**, *109*, 18078-18083, doi:10.1073/pnas.1208835109.
67. Stough, J.M.A.; Yutin, N.; Chaban, Y.V.; Moniruzzaman, M.; Gann, E.R.; Pound, H.L.; Steffen, M.M.; Black, J.N.; Koonin, E.V.; Wilhelm, S.W., et al. Genome and Environmental Activity of a Chrysochromulina parva Virus and Its Virophages. *Frontiers in Microbiology* **2019**, *10*, doi:10.3389/fmicb.2019.00703.
68. Fuhrman, J.A. Marine viruses and their biogeochemical and ecological effects. *Nature* **1999**, *399*, 541-548, doi:10.1038/21119.
69. Liu, H.; Yuan, X.C.; Xu, J.; Harrison, P.J.; He, L.; Yin, K.D. Effects of viruses on bacterial functions under contrasting nutritional conditions for four species of bacteria isolated from Hong Kong waters. *Scientific Reports* **2015**, *5*, doi:10.1038/srep14217.

70. Wilhelm, S.W.; Suttle, C.A. Viruses and Nutrient Cycles in the Sea - Viruses play critical roles in the structure and function of aquatic food webs. *Bioscience* **1999**, *49*, 781-788, doi:10.2307/1313569.
71. Filee, J.; Forterre, P.; Laurent, J. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Research in Microbiology* **2003**, *154*, 237-243, doi:10.1016/s0923-2508(03)00066-4.
72. Sobhy, H. Virophages and Their Interactions with Giant Viruses and Host Cells. *Proteomes* **2018**, *6*, doi:10.3390/proteomes6020023.