2015

Assessment of a Metaviromic Dataset Generated from Nearshore Lake Michigan

Siobhan C. Watkins
Loyola University Chicago, swatkins@luc.edu

Neil Kuehnle
Loyola University Chicago

C Anthony Ruggeri
Loyola University Chicago

Kema Malki
Loyola University Chicago, kmalki@luc.edu

Katherine Bruder
Loyola University Chicago, kbruder@luc.edu

See next page for additional authors

Authors
Siobhan C. Watkins, Neil Kuehnle, C Anthony Ruggeri, Kema Malki, Katherine Bruder, Jinan Elayyan, Kristina Damisch, Naushin Vahora, Paul O'Malley, Brianne Ruggles-Sage, Zachary Romer, and Catherine Putonti

CSIRO PUBLISHING

Marine and Freshwater Research
http://dx.doi.org/10.1071/MF15172

Assessment of a metaviromic dataset generated from nearshore Lake Michigan

Siobhan C. Watkins[A,D], *Neil Kuehnle*[A], *C. Anthony Ruggeri*[A], *Kema Malki*[A], *Katherine Bruder*[A], *Jinan Elayyan*[A], *Kristina Damisch*[A], *Naushin Vahora*[A], *Paul O'Malley*[A], *Brieanne Ruggles-Sage*[A], *Zachary Romer*[B] and *Catherine Putonti*[A,B,C]

[A]Department of Biology, Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL 60660, USA.
[B]Department of Computer Science, Loyola University Chicago, Chicago, IL 60611, USA.
[C]Bioinformatics Program, Loyola University Chicago, Chicago, IL 60660, USA.
[D]Corresponding author. Email: swatkins@luc.edu

Abstract. Bacteriophages are powerful ecosystem engineers. They drive bacterial mortality rates and genetic diversity, and affect microbially mediated biogeochemical processes on a global scale. This has been demonstrated in marine environments; however, phage communities have been less studied in freshwaters, despite representing a potentially more diverse environment. Lake Michigan is one of the largest bodies of freshwater on the planet, yet to date the diversity of its phages has yet to be examined. Here, we present a composite survey of viral ecology in the nearshore waters of Lake Michigan. Sequence analysis was performed using a web server previously used to analyse similar data. Our results revealed a diverse community of DNA phages, largely comprising the order Caudovirales. Within the scope of the current study, the Lake Michigan virome demonstrates a distinct community. Although several phages appeared to hold dominance, further examination highlighted the importance of interrogating metagenomic data at the genome level. We present our study as baseline information for further examination of the ecology of the lake. In the current study we discuss our results and highlight issues of data analysis which may be important for freshwater studies particularly, in light of the complexities associated with examining phage ecology generally.

Received 30 April 2015, accepted 6 August 2015, published online 4 November 2015

Introduction

The study of viruses in the environment, particularly those infectious to bacteria (bacteriophages) and other prokaryotes, has been driven in recent years by several exciting discoveries (Rice *et al.* 2004; Zhao *et al.* 2013). With the advent of improved genomic techniques for molecular-level analyses, microbiologists are slowly uncovering the depth of viral diversity indigenous to areas other than human physiology. Despite the viral renaissance, we still understand very little with regard to how viruses, particularly phages, interact *in situ* with their environment and their hosts.

Phages play a crucial role in the structuring of all microbial communities. They mediate host mortality and drive bacterial genetic diversity and production (Winget *et al.* 2011) in a constant cycle of antagonistic coevolution (Buckling and Rainey 2002). Arguably, one of the most well characterised bacteriophage habitats is the marine (Suttle 2005). This is due largely to the vast impact that marine cyanobacteria have in the oceans in their role as ecosystem engineers (Berman-Frank *et al.* 2003), and the subsequent impact of marine cyanophages upon their ecology. The level of diversity in phage populations in freshwater communities is likely to exceed that found in marine environments, if only owing to the known habitat heterogeneity found in these systems (Maranger and Bird 1995; Wang *et al.* 2012; Berdjeb *et al.* 2013). Generally, freshwater microbial communities have received less attention from microbiologists in comparison to the marine (Ghai *et al.* 2014), especially freshwater phage communities. The investigation of environments that are likely to support highly diverse phage populations is an important step towards identifying key bacteria-phage interactions, such as that between ΦSPM-2 and its cyanobacterial host, which allowed Clokie *et al.* (2006) to identify the virus as being 'photosynthetic' in nature. This will, in turn, inform our understanding of how phages drive important environmental processes.

Metagenomic techniques have become the favoured approach for large-scale examination of microbes in the environment. Although they have inherent limitations (as with all experimental techniques), and should be used to draw conclusions with regard to microbial communities within the scope of strong experimental design, they have allowed for a glimpse into phage diversity and relative abundances for a range of niches, e.g. human microbiota (Reyes *et al.* 2012; Minot *et al.* 2011), soil (Fierer *et al.* 2007), marine environments (Angly *et al.* 2006; Hurwitz and Sullivan 2013) and freshwater environments (Djikeng *et al.* 2009; Roux *et al.* 2012).

Studies of freshwater bacteria suggest the presence of a 'core' group, which is supplemented with 'minor' phyla (Newton *et al.* 2011). Owing to the nature of the host-parasite relationship, we could expect taxonomic classifications of phages to follow that of the bacterial community. However, phage lifestyle and experimental preparation of water samples dictate that this may not be the case. Lysogeny, the capacity of phages to lie dormant within bacterial genomes, has led to a 'mislabelling' problem in the databases serving tools such as BLAST, and the genetic material of prophages is often classified as being bacterial (Casjens 2003). Furthermore, the presence of prophages is not identified by water-sample processing methods, designed to capture viral diversity. For these reasons, phage taxonomy is unlikely to closely follow that of bacteria in the same sample set. Despite this, elucidation of the identifiable features of the accessible phage community in freshwaters remains key to our ultimate understanding of these environments.

In microbial ecology, initial survey data can help support hypothesis-driven experimental analysis, and microbe mining for specific processes (Gijs Kuenen 2008). The data generated forms a foundation for the description of potential metabolic networks and processes, which may subsequently be examined in closer detail. Freshwater systems are a highly dynamic and delicate prospect with regard to bacterial ecology (Poretsky *et al.* 2014) – this level of fluctuation may be scaled up an order of magnitude for phages, considering their known genomic plasticity. The study presented herein describes an initial survey of phages in Lake Michigan, an oligotrophic body of freshwater (Evans *et al.* 2011). To generate a baseline for the examination of viral ecology in Lake Michigan, dsDNA viruses were targeted. The sample collections were processed both as a total composite and as separate viromes, and the current study presents an initial assessment of DNA-based viral diversity in Lake Michigan. Data generated from our metagenomic survey suggest that, based on initial assessment, the general diversity of viruses found in Lake Michigan is, predictably, likely to be specific to the lake itself and highly changeable. We discuss these results in line with the nature of phages as variable entities themselves.

Materials and methods

Sample collection and processing

Two Chicago beaches, Montrose Beach (41°58′0.71″N, 87°38′13.35″W) and 57th Street Beach (41°47′25.54″N, 87°34′41.25″W), were selected as study sites, given their equal proximity from the city centre. No specific permits or permissions were required for the water samples collected from the Chicago Lake Michigan nearshore waters. Sampling was conducted at both sites within recreational swimming areas. Water was collected at a depth of 0.5 m. Four samples (4 L each) were collected from each site, pooled, and processed within 1 h. Sampling was repeated every 10 days from 5 June to 14 August 2013 (16 samples in total).

Viral particles were collected by filtering and concentration. The samples were passed through sterile 0.45-μm cellulose acetate membrane filters (Corning Inc., Corning, NY) and 0.22-μm polyethersulfone membrane filter (MO BIO Laboratories, Carlsbad, CA). Viral particles were then concentrated using a 0.10-μm polypropylene filter (EMD Millipore Corp, Billerica, MA) attached to the Labscale tangential flow filtration (TFF) system (EMD Millipore Corp, Billerica, MA). The filter was cleaned extensively between each sample, using a commercially available Tergazyme preparation.

Viral DNA extraction

Extraction of viral DNA was attempted from the 16 samples using the MO BIO Laboratories PowerWater DNA Isolation Kit (Carlsbad, CA), according to the manufacturer's instructions with one alteration: after optimisation and validation experiments, an additional heat treatment at 70°C for 10 min before initial vortexing was added. It was not possible to extract DNA from all the viral fractions (a common problem encountered during the examination of environmental viruses (Hurwitz and Sullivan 2013)), and nine viral fractions yielded high concentration, high quality DNA. DNA was prepared without any additional amplification steps, therefore favouring the capture of dsDNA viruses (Kim and Bae 2011). To test for bacterial contamination, each extraction underwent PCR testing using primers for the bacterial 16S rRNA gene (Marchesi *et al.* 2001), alongside an *Escherichia coli*-positive control. No amplicons were produced by the viral DNA template.

Sequencing and assembly

Libraries were prepared from the nine samples using the Nextera XT kit for Illumina MiSeq library preparation. The nine samples were multiplexed for a single run of paired-end reads and sequenced by Genewiz (South Plainfield, NJ). Over 15 000 000 raw reads were generated: the raw FASTQ files can be found using NCBI's BioProject database (Accession: PRJNA248239).

Paired-end assembly and contig assembly were performed for each sample using MetaVelvet (Namiki *et al.* 2012) with a hash length of $k = 29$. The numbers of contigs and paired-end reads generated for each sample are detailed in Table 1.

Taxonomy classification and BLAST-based comparative viromics

Taxonomic classification was conducted on the nine samples using the web server Metavir (Roux *et al.* 2011). Owing to input restrictions of Metavir, the longest 300 000 paired reads (Table 1) were selected for analysis for each of the nine. Metavir assesses virome composition using the GAAS tool (Angly *et al.* 2009). BLAST analyses by Metavir were performed using the RefSeq complete viral genomes protein sequences database (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/, accessed 27 August 2015); the release of the database used for comparison was that of 18 January 2014 and contained 173 122 protein sequences. The e-value cut off was 10e-5. 65% of the data generated by our study and processed through Metavir was determined to have originated from phage genomes. In addition to taxonomic classification, Metavir was used to compute rarefaction curves (Fig. S1 of the Supplementary material).

Table 1. Summary statistics of sequencing and assembly for each virome

Virome label	Collection site, date	Number of paired-end reads	Number of assembled contigs	Contig lengths used in Metavir
A	57th Street, 5 June 2013	2 202 471	722 253	345–500
B	Montrose, 5 June 2013	1 123 802	565 244	317–500
C	Montrose, 15 June 2013	1 380 226	808 040	290–500
D	57th Street, 25 June 2013	1 626 554	81 777	363–500
E	Montrose, 25 June 2013	1 547 157	1 049 321	377–500
F	Montrose, 5 July 2013	1 854 527	1 199 969	371–500
G	57th Street, 15 July 2013	2 166 575	948 441	359–500
H	Montrose, 25 July 2013	2 351 827	1 428 015	330–500
I	57th Street, 14 August 2013	1 415 911	648 250	296–500

Comparative analysis of viromes based on contig assembly

Comparisons to publicly available freshwater viromes were performed in Metavir by performing tBLASTx comparisons using a subsample of the sequences taken from a virome, after data normalisation. Each of the nine viromes were separately assembled in a pair-wise fashion: reads generated from one sample were assembled to the reads generated from another sample, using MetaVelvet. This cross-sample assembly strategy can identify any similarities in the same taxa present within two samples.

ORF generation

The 20 longest contigs assembled by MetaVelvet for each of the nine samples were examined further. Open reading frames (ORFs) within each contig were predicted using GeneMarkS (Besemer *et al.* 2001). Each predicted ORF was then BLASTed against all records in the non-redundant (nr) protein sequences database, using the blastx algorithm from the NCBI web interface (results not shown). BLAST homologies having an e-value less than 10e-3 were considered putative hits.

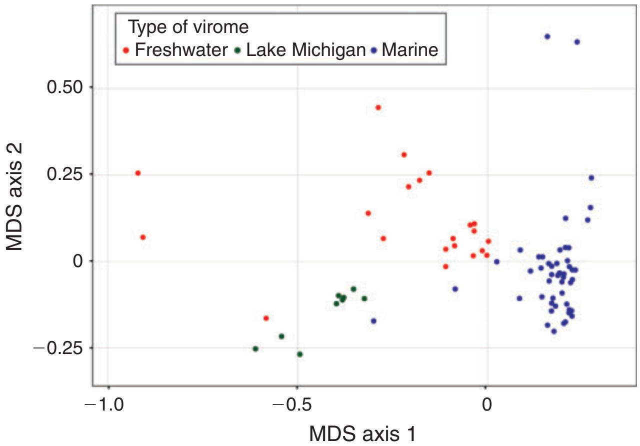

Fig. 1. MDS plots of freshwater (red) and marine (blue) publicly available (Metavir) viromes with the nine Lake Michigan viromes (green). Bray–Curtis dissimilarity matrices of BLAST hits from metaviromic data were calculated in Metavir and used to represent the relative distances between individual samples (stress value 9%).

Results

Overview of Lake Michigan viromes

Nine viromes were generated from samples collected from the nearshore waters of two Chicago area beaches, Montrose Beach and 57th Street Beach, and analysed to assess viral diversity. Samples were evaluated without reference to spatial or temporal difference, and were ultimately pooled in order to assess the range of diversity observed (i.e. as a composite group). Isolated DNA produced $>$15 000 000 raw reads (see Methods).

Rarefaction curves (Fig. S1) generated from the raw data suggested that total viral diversity was captured in only one of the nine viromes. As with most studies examining viral metagenomics, most of raw reads from Lake Michigan did not show any significant sequence similarity to current viral data: only 6.9% were found to produce a BLAST score of $\geq$50 against entries in the NCBI RefSeq viral protein database.

Comparison of Lake Michigan viromes

The Lake Michigan viromes were compared directly to 20 viromes generated from freshwater environments (publicly available) using the Metavir virome-comparison function which is based on the BLAST results for each sample (data normalisation is performed, by genome length, as part of the Metavir package (Roux *et al.* 2011)). The Lake Michigan samples clustered together, largely distinct from the other Metavir viromes (Fig. 1) but demonstrated very slight overlap with both freshwater and marine environments.

Additionally, the complete set of contigs generated for each of the nine Lake Michigan viromes was assembled in a pair-wise manner using the assembler MetaVelvet. This cross-sample assembly strategy can identify similarities in taxa present within two samples, for example, whether representative of viral species or coding sequences present within both samples. The contigs produced by this assembly were used to compare the nine viromes to one-another. These analyses suggested a considerable level of diversity within the composite sample, alongside a core element (greens and yellow, associated with red) (Fig. 2).

Taxonomic classification of viromes

The BLAST results (against RefSeq) for the composite virome, as expected, largely comprised members of the Caudovirales comprising the families *Myoviridae*, *Siphoviridae* and

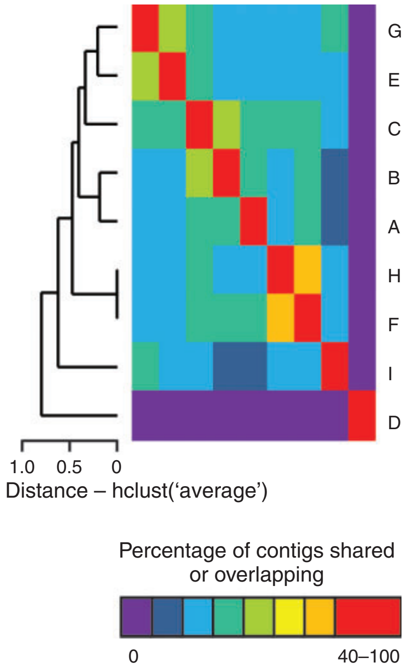

Fig. 2. Diversity within the composite Lake Michigan virome (based on the nine original samples), as determined by cross-sample assembly in MetaVelvet.

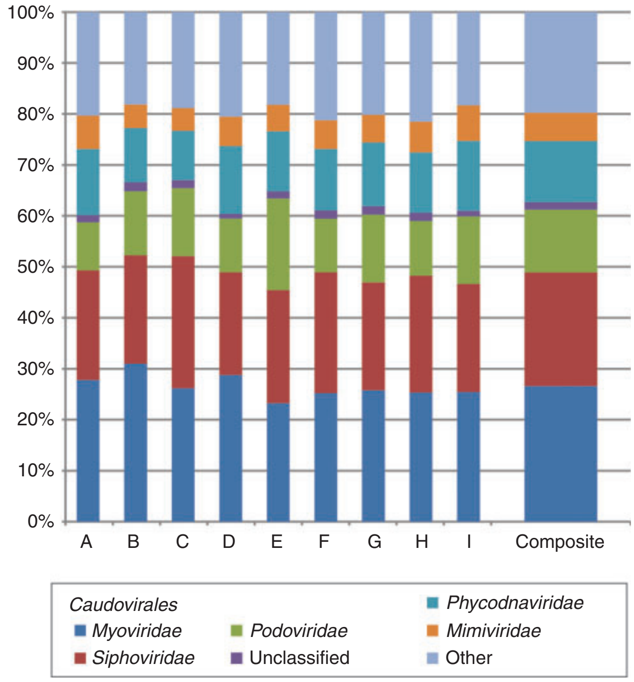

Fig. 3. BLAST-based taxonomic classification (based on matches to the RefSeq database) of the nine Lake Michigan viromes separately and as a composite. Classifications relate to positive (i.e. known) hits only, and do not include hypothetical matches.

Podoviridae (Fig. 3). Although most of the hits were to these dsDNA phages, similarities to ssDNA phages as well as eukaryotic-infecting viruses were also observed (represented by the group 'other'; Fig. 3). The Krona tool (Ondov *et al.* 2011), available as part of Metavir, was used to visualise matches between the composite Lake Michigan sequences and viral proteins in the RefSeq database (Fig. 4). The composite virome shows 95% of the protein hits are to coding regions within dsDNA (no RNA stage) viruses: 65% of which were for Caudovirales.

Viral-associated proteins

Via Metavir, BLAST (against RefSeq) returned matches to proteins belonging to a variety of viral species. Some phage species (1182) were represented throughout the composite virome (i.e. all nine samples). The majority (80%) of these hits pertaining to these phages were to several in particular. Numerous hits to the *Planktothrix* phage PaV-LD (NC_016564) were observed throughout the nine viromes and the composite; Fig. 4 shows that this phage was determined to comprise 4% of the entire virome. Further investigation showed that nearly all of these hits; however, (over 7000; Fig. 5*a*) were to the gene PaVLD_ORF033R. This protein is annotated as an ABC-transporter protein (Gao *et al.* 2012) in RefSeq. Fig. 5*a* also highlights two other genes specific to this phage, present in comparatively low frequency: a DNA helicase and a peptidase.

Further investigation uncovered that all nine separate viromes generated numerous hits to phages infectious to *Burkholderia*: 50% of the 72 annotated protein-coding genes within the 47 399 bp genome of the *Burkholderia* phage BcepB1A (NC_005886) (Summer *et al.* 2006) were found within the Lake Michigan virome. Most of the hits were from three of our original samples (Fig. 5*b*); however, numerous matches to other phages infecting *Burkholderia cepacia* were also present throughout the composite. High coverage for species of phage known to possess larger genomes was also observed, for example, the 252 401 bp genome of *Prochlorococcus* phage P-SSM2 (NC_006883) shown in Fig. 5*c*. Unlike *Planktothrix* phage PaV-LD, a broader distribution of the genomes of the *Burkholderia* and *Prochlorococcus* phages was present throughout the sample set.

BLAST matches to eukaryote-infecting viruses were minimal, the exception being the genomes of the giant amoeba-infecting viruses. Each individual Lake Michigan virome exhibited similarities to hundreds of different coding regions within the genomes of the *Mimiviridae*, *Moumouviridae*, *Megaviridae* and *Pandoraviridae*. For instance, over 16% of the annotated genes in *Acanthamoeba polyphaga mimivirus* (NC_014649) were found in the composite virome (Fig. 5*d*).

Comparing sequence composition of contigs to existing annotated metagenomic information can also give us insight into the potential function of viruses. The BLAST hits to phage species produced by the Metavir analyses for the nine LM viromes were aggregated by function based upon their annotated protein product (Table 2). The number of BLAST hits per annotated function varied considerably between the nine viromes (a range of 91–635). Functions with the greatest number of hits were further investigated. As is typical of similar datasets, most of the generated ORFs were assigned to hypothetical proteins.

Fig. 4. Krona chart representing taxonomic classifications identified using BLAST searches between the composite Lake Michigan virome contigs and RefSeq viral proteins. The interactive Krona charts for all of the Lake Michigan viromes are available at http://metavir-meb.univ-bpclermont.fr/, accessed 25 August 2015.

Discussion

Environmental phages are one of the most abundant, yet understudied elements of microbial ecology. Our level of knowledge regarding the total diversity and genetic composition of environmental phages is minimal. Evidence suggests that communities of phages found in freshwaters are likely to represent a more complex cohort than that of their marine counterparts (Kimura *et al.* 2013; Baker *et al.* 2006), in congruence with the more heterogeneous nature of freshwater. Broadly speaking, in comparison to marine communities, freshwaters are understudied regardless, and so the scope of this expected greater diversity is unknown. Samples were collected from two Chicago beaches of Lake Michigan, and the dsDNA-based viral diversity was examined, with whole genome sequencing. Nine samples were processed and assessed both separately and as a composite.

The nearshore waters of a large lake differ in composition, physicochemically, from offshore waters (Yurista *et al.* 2012; Pilcher *et al.* 2015), so it follows that bacterial and phage community composition will vary from position to position and over time: on both large and small scales. However, capturing and successfully interrogating the changes in phage diversity temporally and spatially would require an exhaustive sampling and deep-sequencing effort. Instead, we examined our samples as a composite, in order to generate a 'baseline' representation of the phage community present in our samples from Lake Michigan.

For the purposes of assessment of diversity in the samples collected, the current study was designed to assess the presence of DNA viruses, in congruence with most of the other metaviromics studies performed in the environment. Furthermore, the web server we used to process our data (Metavir) has

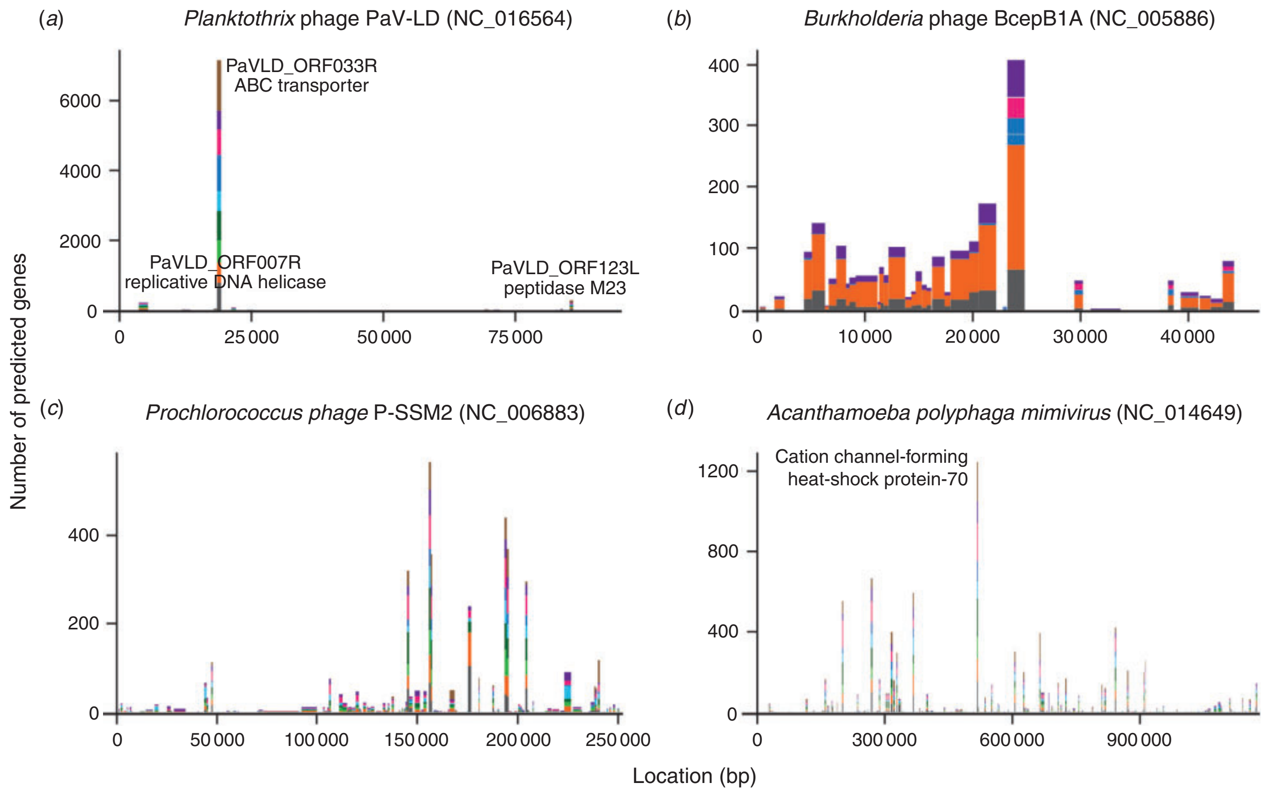

Fig. 5. Highly abundant viral proteins present in the Lake Michigan virome. The *x*-axis represents the genomes of one of four viruses: *Planktothrix* phage PaV-LD (*a*), *Burkholderia* phage BcepB1A (*b*), *Prochlorococcus* phage P-SSM2 (*c*) and *Acanthamoeba polyphaga mimivirus* (*d*). Colour coding identifies the original sample (of the nine separate viromes) of origin for each protein.

Table 2. Annotations of proteins most frequently observed within the Lake Michigan virome contigs processed through Metavir

nr, non-redundant

Annotated function of hit	Number of BLAST hits (nr) to Lake Michigan viromes (percentage of all hits)
Hypothetical or unnamed protein	194 155 (15.29%)
DNA polymerase	69 175 (5.45%)
Terminase, large subunit	45 495 (3.58%)
DNA helicase	23 555 (1.85%)
DNA integrase	22 728 (1.79%)
Portal protein	21 979 (1.73%)
Structural protein	14 534 (1.14%)

previously been used with marine, brackish and freshwater sample sets (Roux *et al.* 2012). Owing to the nature of sample preparation (for example, multiple displacement amplification is known to produce a bias for ssDNA viruses (Kim and Bae 2011)), we theorised that most of BLAST hits returned by our sequencing effort would likely belong to dsDNA viruses. As anticipated, dsDNA phages were found to mostly comprise viruses in all datasets. Direct comparison of the Lake Michigan virome generated in the current study with similar research presents inherent challenges, owing to the innate levels of heterogeneity, which each habitat is likely to exhibit, as well as differences in sample preparation. Despite these challenges, we were able to compare the LM virome with other freshwater viromes using BLAST-based analyses using Metavir. The comparison revealed that the dataset generated from the current study appears to be divergent from previous studies (Fig. 1). In light of known vacillations in community composition in freshwaters this is to be expected, and suggests that disparity between freshwater samples from different locations, as well as between freshwater and marine samples, is also likely. This said, rarefaction curves produced as part of Metavir's analysis (Fig. S1) suggest that out of the nine viromes produced, only one captured the entire representative diversity. Metavir restricts inputs to 300 000 reads, and Roux *et al.* (2012) described similar findings for two freshwater lake samples: viral diversity was not entirely captured according to rarefaction curves. In addition, the potential for bias in sampling, sample preparation and sequencing is considerable (Schoenfeld *et al.* 2010), and therefore our low diversity virome may have fallen subject to these issues. However, quality of this sample was assessed before sequencing (as part of standard QC protocols before Illumina sequencing) and was judged to be congruent with the others. Owing to the general paucity of knowledge in this area, it is difficult to interpret why one sample may have been so markedly different to others taken from the same environment, albeit at different times. However, in freshwaters such differences in diversity, on a temporal and spatial basis, are possible, if not expected. These discrepancies are counteracted to some extent by our decision to analyse our samples as a composite. Overall, however, these results suggest a considerable reservoir of unexplored viral ecology.

Furthermore, the BLAST-based comparisons highlighted only a small fraction of the genomic content within the Lake Michigan viromes: this is largely unavoidable owing to the paucity of viral data in RefSeq, for example, which means most of the hits are of unknown origin. Regardless of the environmental niche investigated, most of the sequence reads generated from projects examining metaviromes in particular, demonstrate infrequent homology to previously annotated function or taxonomy (Yin and Fischer 2008). This is owing to the low frequency of entire phage genomes in the RefSeq database and, in turn, an even lower incidence of entirely and accurately annotated phage genomes. This remains a key confounding issue with regard to examining phage diversity in the environment and further highlights the requirement for continued culture-based investigations based on physical isolation of phages.

Taxonomically, most of the viruses in our virome belong to the order Caudovirales. Caudovirales, dsDNA viruses comprising the families *Myoviridae*, *Podoviridae* and *Siphoviridae*, and represent the low-hanging fruit of the phage world. They are easy to isolate from environmental samples and maintain in the laboratory, and therefore they are easy to find in metaviromic datasets. For this reason they have substantial presence in the RefSeq database and the International Committee on Taxonomy of Viruses database (http://ictvonline.org/virustaxonomy.asp, accessed 8 August 2015). We identified several phage taxa of interest to examine further. Usually, phages are named for their bacterial hosts; however, this can represent a challenge when classifying diversity representatively. With a BLAST e-value cut-off of 10e-5, all nine viromes showed numerous hits to the *Planktothrix* phage PaV-LD (NC_016564). Further investigation showed that nearly all of these matches (Fig. 5*a*) were to the gene PaVLD_ORF033R, which is annotated as an ABC-transporter protein (Gao *et al.* 2012). The presumptive abundance of this gene in our samples is extremely unlikely to represent a single 'species' of phage only, and is more likely owing to the high presence of non-species or virus-specific ABC-transporter proteins (ubiquitous throughout all extant prokaryotic and eukaryotic phyla alike) in the sample. Owing to previous identification of a high incidence of *Burkholderia* in samples taken from the lake (data not shown), *Burkholderia* phage BcepB1A was highlighted for further examination.

In addition to hits to the ABC-transporter coding region (YP_004957 306) for the *Planktothrix* phage PaV-LD genome (Fig. 5*a*), hits to other phage species, including *Bacillus* phage SPbeta and *Bacillus* phage G, were also identified. Further investigation into these hits revealed that sequence similarity between the Lake Michigan virome contigs and the RefSeq database was localised to the ABCC_MRP_Like domain (CDD ID: cd03228). Expanding our analysis to all hits, not just to phage species, there were a significant number that were analogous to coding regions annotated as ABC transporters within the genomes of amoeba-infecting viruses. As the Metavir analysis is limited to the RefSeq annotated viral-genome collection, we checked to see if other viral metagenomic studies had also observed this function in their sample. Selecting the protein sequence for this domain from the *Bacillus* phage-SPbeta genome, BLASTp searches were conducted against the non-redundant protein-sequences database (limiting the search to sequences taxonomically classified as viruses) as well as the metagenomic proteins database (env_nr). In the case of the former, numerous hits to sequences of the taxon *Phycodnaviridae* were identified. This suggests that there is incongruity between the initial classification of *Bacillus* phage and the second search, which employed more stringent parameters. Similarly, the BLASTp search to the env_nr database identified statistically significant hits (E-value = 8e-37) to sequences generated from global-ocean sampling studies. Although *Prochlorococcus* is a marine cyanobacterium, it has been determined previously that P-SSM2, a myovirus, is likely to contain several core genes, both belonging to T4-like groups and those which are cyanobacterial in origin (Sullivan *et al.* 2005). *Prochlorococcus* is a dominant genus of cyanobacteria in the marine environment, but a vast catalogue of freshwater cyanobacteria will be indigenous and present in association with freshwater cyanophages. Cyanobacteria, and by definition, cyanophages, contain highly mobile genetic elements and therefore both groups, both marine and freshwater may possess the genes identified in our sample set (Ivanikova *et al.* 2007). This explains why genes that are, ostensibly, marine in origin, may have validity in a freshwater dataset. Our dataset contained hits against P-SSM2 (Fig. 5*c*): >50% of its annotated coding regions. This phage has a significantly larger genome than that of the *Burkholderia* phage, that of the mimivirus is larger still: based on the distribution of hits to these genomes, we hypothesise that relatives of these phages are actually present in the lake, as opposed to PaV-LD. Hits to the *Plankthothrix* phage, in this case, likely reveal only the presence of ABC-transporter genes. These analyses reconfirm that in a dataset as complex as an aquatic virome, where most of BLAST hits are unknown, the presence of specific genes is what may give insight into community composition, as opposed to the species assigned to that gene specifically.

Those ORFs that returned more descriptive hits (as opposed to hypothetical, and including DNA helicases, DNA integrases, DNA polymerases, portal proteins, structural proteins and the large subunit of terminases), all belonged to proteins most prevalent in single-annotated phage genomes (Kristensen *et al.* 2013) as opposed to those mined from metagenomic data. Integrases are highly ubiquitous, and can be shared by phages, whereas previous bioinformatic analysis (Kristensen *et al.* 2013), found that many other functions are phage-specific (i.e. pertaining to phages, and not bacteria), including the large subunit of terminases, portal proteins, and structural proteins. Other phage-specific ORFs, which are frequently found in annotated viral genomes, such as those belonging to the cro or cI-repressor system (important in lytic and lysogenic development (Folkmanis *et al.* 1977)) and homing HNH endonucleases (important in lateral transfer (Friedrich *et al.* 2007)), were also identified in the Metavir hits, albeit far fewer in number (1315 and 3855 respectively). This suggests the presence of lysogenic or potentially lysogenic phages in the Lake Michigan viromes.

Phages are able to propagate only by the successful infection of a host. There are two major mechanisms for infection: (1) lysis, in which viral progeny are released from the host cell as a fatal event, and (2) lysogeny, during which genomic materials from a phage are integrated with that of a host cell. Over time, this may result in genotypic and phenotypic changes to the host cell, or eventual cell death. Therefore, most genomic information available for phages pertains to lytic phages. Of course, phages may also change lifestyle over time, a factor which may only be determined in culture (Watkins *et al.* 2014). Clearly, the two lifestyles may confound metagenomic survey-based examinations of total viral diversity, as sample preparation methods are largely designed to capture 'free' viruses. In addition, one phage may infect more than one host, and in some cases, this broad host-range may extend across bacterial genera. Phages previously examined and labelled as infectious to one genus, may ultimately have a different preferred host. This phenomenon, may, similarly, be changeable according to resource limitation or abundance, physical factors, or the proliferation of a specific taxon of bacteria over another. These factors, among others, should be taken into consideration while examining metagenomics datasets.

The nearshore waters of Lake Michigan represent a viral population, which is diverse from other publicly available datasets and which demonstrates phylogenetic divergence from previously studied viral communities. Although the viromes analysed in the current study are composed largely of groups of dsDNA phages commonly present in other such datasets (as expected), they are a potentially rich source of further information regarding the genetic heterogeneity of freshwater systems in general. The data obtained from the current study, from a composite sampling regime, has been used to assess the diversity of Lake Michigan by examining the presence of commonly seen viral-associated genes and proteins. Not only will this work enable continued investigations into the viral ecology of freshwater systems, including studies designed to infer impact and interaction of phage communities on and with their hosts, but it has laid a foundation for the mammoth task of examining the entire community of Lake Michigan.

Acknowledgements

This research was funded by the NSF (1 149 387) (CP). The authors would like to thank Jourdan Howard, Alex Kula, Matthew Putonti, Daniel Searle, and Nick White for their assistance in collecting samples. Thanks also to Mr. Thomas Hatzopoulos for his assistance in collecting data.

References

Angly, F., Felts, B., Breitbart, M., Salamon, P., Edwards, R., Carlson, C., Chan, A., Haynes, M., Kelley, S., Liu, H., Mahaffy, J., Mueller, J., Nulton, J., Olson, R., Parsons, R., Rayhawk, S., Suttle, C., and Rohwer, F. (2006). The marine viromes of four oceanic regions. *PLoS Biology* **4**, e368. doi:10.1371/JOURNAL.PBIO.0040368

Angly, F., Willner, D., Prieto-Davó, A., Edwards, R., Schmieder, R., Vega-Thurber, R., Antonopoulos, D., Barott, K., Cottrell, M., Desnues, C., Dinsdale, E., Furlan, M., Haynes, M., Henn, M., Hu, Y., Kirchman, D., McDole, T., McPherson, J., Meyer, F., Miller, R., Mundt, E., Naviaux, R., Rodriguez-Mueller, B., Stevens, R., Wegley, L., Zhang, L., Zhu, B., and Rohwer, F. (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Computational Biology* **5**, e1000593. doi:10.1371/JOURNAL.PCBI.1000593

Baker, A., Goddard, V., Davy, J., Schroeder, D., Adams, D., and Wilson, W. (2006). Identification of a diagnostic marker to detect freshwater cyanophages of filamentous cyanobacteria. *Applied and Environmental Microbiology* **72**, 5713–5719. doi:10.1128/AEM.00270-06

Berdjeb, L., Pollet, T., Chardon, C., and Jacquet, S. (2013). Spatio-temporal changes in the structure of archaeal communities in two deep freshwater lakes. *FEMS Microbiology Ecology* **86**, 215–230. doi:10.1111/1574-6941.12154

Berman-Frank, I., Lundgren, P., and Falkowski, P. (2003). Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Research in Microbiology* **154**, 157–164. doi:10.1016/S0923-2508(03)00029-9

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research* **29**, 2607–2618. doi:10.1093/NAR/29.12.2607

Buckling, A., and Rainey, P. B. (2002). Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society of London B: Biological Sciences* **269**, 931–936. doi:10.1098/RSPB.2001.1945

Casjens, S. (2003). Prophages and bacterial genomics: what have we learned so far? *Molecular Microbiology* **49**, 277–300. doi:10.1046/J.1365-2958.2003.03580.X

Clokie, M., Shan, J., Bailey, S., Jia, Y., Krisch, H., West, S., and Mann, N. (2006). Transcription of a 'photosynthetic' T4-type phage during infection of a marine cyanobacterium. *Environmental Microbiology* **8**, 827–835. doi:10.1111/J.1462-2920.2005.00969.X

Djikeng, A., Kuzmickas, R., Anderson, N., and Spiro, D. (2009). Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* **4**, e7264. doi:10.1371/JOURNAL.PONE.0007264

Evans, M. A., Fahnenstiel, G., and Scavia, D. (2011). Incidental oligotrophication of North American Great Lakes. *Environmental Science & Technology* **45**, 3297–3303. doi:10.1021/ES103892W

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., Robeson, M., Edwards, R. A., Felts, B., Rayhawk, S., Knight, R., Rohwer, F., and Jackson, R. B. (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology* **73**, 7059–7066. doi:10.1128/AEM.00358-07

Folkmanis, A., Maltzmann, W., Mellon, P., Skalka, A., and Echols, H. (1977). The essential role of the cro gene in lytic development by bacteriophage ʎ. *Virology* **81**, 352–362. doi:10.1016/0042-6822(77)90151-9

Friedrich, N., Torrents, E., Gibb, E. A., Sahlin, M., Sjöberg, B.-M., and Edgell, D. R. (2007). Insertion of a homing endonuclease creates a genes-in-pieces ribonucleotide reductase that retains function. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 6176–6181. doi:10.1073/PNAS.0609915104

Gao, E.-B., Gui, J.-F., and Zhang, Q.-Y. (2012). A novel cyanophage with a cyanobacterial nonbleaching protein A gene in the genome. *Journal of Virology* **86**, 236–245. doi:10.1128/JVI.06282-11

Ghai, R., Mizuno, C., Picazo, A., Camacho, A., and Rodriguez-Valera, F. (2014). Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. *Molecular Ecology* **23**, 6073–6090. doi:10.1111/MEC.12985

Gijs Kuenen, J. (2008). Anammox bacteria: from discovery to application. *Nature* **6**, 320–326.

Hurwitz, B., and Sullivan, M. (2013). The Pacific Ocean virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One* **8**, e57355. doi:10.1371/JOURNAL.PONE.0057355

Ivanikova, N., Popels, L., McKay, R., and Bullerjahn, G. (2007). Lake Superior supports novel clusters of cyanobacterial picoplankton. *Applied and Environmental Microbiology* **73**, 4055–4065. doi:10.1128/AEM.00214-07

Kim, K.-H., and Bae, J.-W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded

DNA viruses. *Applied and Environmental Microbiology* **77**, 7663–7668. doi:10.1128/AEM.00289-11

Kimura, S., Sako, Y., and Yoshida, T. (2013). Rapid *Microcystis cyanophage* gene diversification revealed by long- and short-term genetic analyses of the tail sheath gene in a natural pond. *Applied and Environmental Microbiology* **79**, 2789–2795. doi:10.1128/AEM.03751-12

Kristensen, D., Waller, A., Yamada, T., Bork, P., Mushegian, A., and Koonin, E. (2013). Orthologous gene clusters and taxon signature genes for viruses of prokaryotes. *Journal of Bacteriology* **195**, 941–950. doi:10.1128/JB.01801-12

Maranger, R., and Bird, D. (1995). Viral abundance in aquatic systems: a comparison between marine and fresh waters. *Marine Ecology Progress Series* **121**, 217–226. doi:10.3354/MEPS121217

Marchesi, J., Weightman, A., Cragg, B., Parkes, R., and Fry, J. (2001). Methanogen and bacterial diversity and distribution in deep gas hydrate sediments from the Cascadia Margin as revealed by 16S rRNA molecular analysis. *FEMS Microbiology Ecology* **34**, 221–228. doi:10.1111/J.1574-6941.2001.TB00773.X

Minot, S., Sinha, R., Chen, J., Hongzhe, Li., Keilbaugh, S. A., Wu, G. D., Lewis, J. D., and Bushman, F. D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Research* **21**, 1616–1625. doi:10.1101/GR.122705.111

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research* **40**, e155. doi:10.1093/NAR/GKS678

Newton, R., Jones, S., Eiler, A., McMahon, K., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiology and Molecular Biology Reviews* **75**, 14–49. doi:10.1128/MMBR.00028-10

Ondov, B., Bergman, N., and Phillippy, A. (2011). Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12**, 385–394. doi:10.1186/1471-2105-12-385

Pilcher, D. J., McKinley, G. A., Bootsma, H. A., and Bennington, V. (2015). Physical and biogeochemical mechanisms of internal carbon cycling in Lake Michigan. *Journal of Geophysical Research* **120**, 2112–2128.

Poretsky, R., Rodriguez-R, L., Luo, C., Tsementzi, D., and Konstantinidis, K. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS One* **9**, e93827. doi:10.1371/JOURNAL.PONE.0093827

Reyes, A., Semenkovich, N., Whiteson, K., Rohwer, F., and Gordon, J. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature Reviews. Microbiology* **10**, 607–617. doi:10.1038/NRMICRO2853

Rice, G., Tang, L., Stedman, K., Roberto, F., Spuhler, J., Gillitzer, E., Johnson, J., Douglas, T., and Young, M. (2004). The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7716–7720. doi:10.1073/PNAS.0401773101

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., and Enault, F. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27**, 3074–3075. doi:10.1093/BIOINFORMATICS/BTR519

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., Sime-Ngando, T., and Debroas, D. (2012). Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* **7**, e33641. doi:10.1371/JOURNAL.PONE.0033641

Schoenfeld, T., Liles, M., Wommack, K., Polson, S., Godiska, R., and Mead, D. (2010). Functional viral metagenomics and the next generation of molecular tools. *Trends in Microbiology* **18**, 20–29. doi:10.1016/J.TIM.2009.10.001

Sullivan, M., Coleman, M., Weigele, P., Rohwer, F., and Chisholm, S. (2005). Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biology* **3**, e144. doi:10.1371/JOURNAL.PBIO.0030144

Summer, E., Gonzalez, C., Bomer, M., Carlile, T., Embry, A., Kucherka, A., Lee, J., Mebane, L., Morrison, W., Mark, L., King, M., LiPuma, J., Vidaver, A., and Young, R. (2006). Divergence and mosaicism among virulent soil phages of the *Burkholderia cepacia* complex. *Journal of Bacteriology* **188**, 255–268. doi:10.1128/JB.188.1.255-268.2006

Suttle, C. (2005). Viruses in the sea. *Nature* **437**, 356–361. doi:10.1038/NATURE04160

Wang, Y., Sheng, H.-F., He, Y., Wu, J.-Y., Jiang, Y.-X., Tam, N., and Zhou, H.-W. (2012). Comparison of the levels of bacterial diversity in freshwater, intertidal wetland, and marine sediments by using millions of illumina tags. *Applied and Environmental Microbiology* **78**, 8264–8271. doi:10.1128/AEM.01821-12

Watkins, S., Smith, J., Hayes, P., and Watts, J. (2014). Characterisation of host growth after infection with a broad-range freshwater cyanopodophage. *PLoS One* **9**, e87339. doi:10.1371/JOURNAL.PONE.0087339

Winget, D., Helton, R., Williamson, K., Bench, S., Williamson, S., and Wommack, K. (2011). Repeating patterns of virioplankton production within an estuarine ecosystem. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 11 506–11 511. doi:10.1073/PNAS.1101907108

Yin, Y., and Fischer, D. (2008). Identification and investigation of ORFans in the viral world. *BMC Genomics* **9**, 24. doi:10.1186/1471-2164-9-24

Yurista, P. M., Kelly, J. R., Miller, S. E., and Van Alstine, J. D. (2012). Water quality and plankton in the United States nearshore waters of Lake Huron. *Environmental Management* **50**, 664–678. doi:10.1007/S00267-012-9902-X

Zhao, Y., Temperton, B., Thrash, J., Schwalbach, M., Vergin, K., Landry, Z., Ellisman, M., Deerinck, T., Sullivan, M., and Giovannoni, S. (2013). Abundant SAR11 viruses in the ocean. *Nature* **494**, 357–360. doi:10.1038/NATURE11921

Marine and Freshwater Research
http://dx.doi.org/10.1071/MF15172_AC

Supplementary material

Assessment of a metaviromic dataset generated from nearshore Lake Michigan

Siobhan C. Watkins[A,D], *Neil Kuehnle*[A], *C. Anthony Ruggeri*[A], *Kema Malki*[A], *Katherine Bruder*[A], *Jinan Elayyan*[A], *Kristina Damisch*[A], *Naushin Vahora*[A], *Paul O'Malley*[A], *Brieanne Ruggles-Sage*[A], *Zachary Romer*[B] *and Catherine Putonti*[A,B,C]

[A]Department of Biology, Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL 60660, USA.

[B]Department of Computer Science, Loyola University Chicago, Chicago, IL 60611, USA.

[C]Bioinformatics Program, Loyola University Chicago, Chicago, IL 60660, USA.

[D]Corresponding author. Email: swatkins@luc.edu

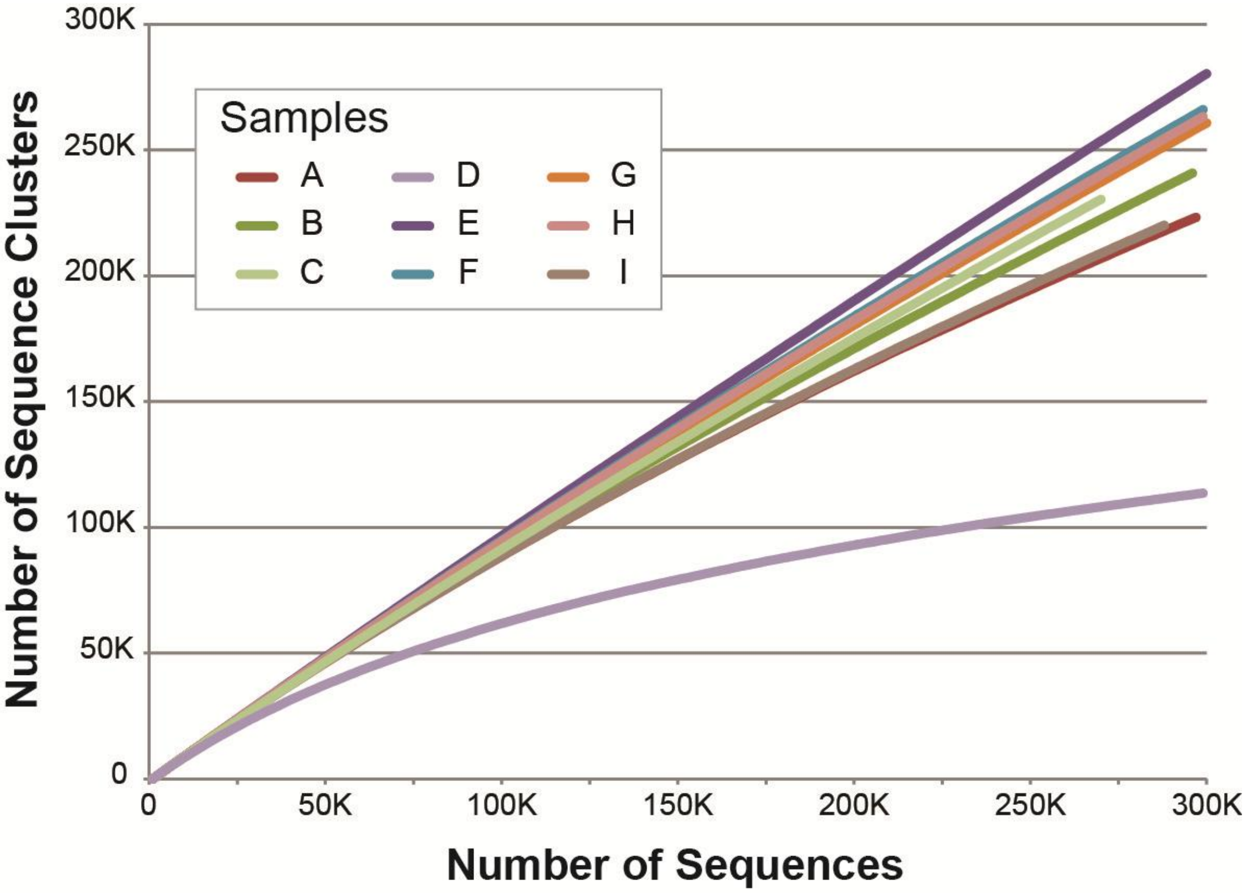

Fig. S1. Rarefaction curves for the Lake Michigan viromes, as generated in Metavir.