**Trying to find many needles in lots of haystacks: preliminary investigations into freshwater bacterial-viral interactions in a divergent sample set from Lake Michigan**

Siobhan C.Watkins1\*, Neil Kuehnle1, C. Anthony Ruggeri1, Kema Malki1, Katie Bruder1, Jinan Elayyan1, Kristina Damisch1, Naushin Vahora1, Paul O’Malley1, Brieanne Ruggles-Sage1, Zachary Romer2, Catherine Putonti1,2,3.

1 Department of Biology, Loyola University Chicago, Chicago, IL USA 60660

2 Department of Computer Science, Loyola University Chicago, Chicago, IL USA 60611

3 Bioinformatics Program, Loyola University Chicago, Chicago, IL USA 60660

**SCW:** [**swatkins@luc.edu**](mailto:swatkins@luc.edu)**; NK:** [**nkuehnle@luc.edu**](mailto:nkuehnle@luc.edu)**; CAR:** [**cruggeri@luc.edu**](mailto:cruggeri@luc.edu)**; KM:** [**kmalki@luc.edu**](mailto:kmalki@luc.edu)**; KB:** [**kbruder@luc.edu**](mailto:kbruder@luc.edu)**; JE:** [**jelayyan@luc.edu**](mailto:jelayyan@luc.edu)**; KD:** [**kdamisch@luc.edu**](mailto:kdamisch@luc.edu)**; NV:** [**nvahora1@luc.edu**](mailto:nvahora1@luc.edu)**; PO:** [**pomalley1@luc.edu**](mailto:pomalley1@luc.edu)**; BR-S:** [**brugglessage@luc.edu**](mailto:brugglessage@luc.edu)**; ZR:** [**zromer@luc.edu**](mailto:zromer@luc.edu)**; CP: cputonti@luc.edu**

\*Corresponding author, email: swatkins@luc.edu

**Abstract**

Bacteriophages are known to influence bacterial diversity, and therefore key environmental processes, in environmental niches across the globe. In the oceans particularly, microbiologists are beginning to focus highly resolved strategies on the assessment of viral diversity and host-viral interactions. The freshwater cohort is understudied in comparison, but the level of diversity present in all other aspect of freshwater biology and water chemistry suggest that viral diversity is likely to be considerably more complex than that observed in marine niches. Few studies have comprehensively assessed the bacterial and viral component of freshwater metagenomes, however, the comprehensive assessment of such heterogeneous habitats is an important step in furthering our general knowledge of how viral populations shape bacterial activity in the environment. This study presents results obtained by sequencing the bacterial and viral component of samples collected during assessment of temporal and spatial variability along nearshore Lake Michigan, Chicago. Our study has allowed us to draw initial conclusions regarding the rich diversity of viral diversity present in these areas, despite the paucity of environmental viral sequences present in databases such as BLAST. Through grouping and assessing genomic information on a phylogenetic and functional level we begin to discuss the highly complex relationship between the two groups, and how they may reflect freshwater environments generally. To our knowledge this is the first such study performed on Lake Michigan.

Keywords: freshwater, bacteriophages, bacteria, whole genome sequencing, metagenomics, diversity, host-viral interactions

**Introduction**

The study of viruses found in the environment, particularly those infectious to prokaryotes, has been driven in very recent years by a number of exciting discoveries, for example, the existence of extremely large viruses [1], and viruses which are, in turn, capable of infecting them [2]. With the advent of improved genomic techniques for viral discovery, microbiologists are beginning to slowly uncover the depth of viral diversity indigenous to the planet [3]. Despite the viral renaissance, we still understand very little with regard to how viruses and bacteriophages (viruses that infect bacteria) interact *in situ* with their hosts in the environment.

Bacteriophages, or phages, play a crucial role in the structuring of microbial communities, generally: they mediate host mortality on a large scale and drive bacterial genetic diversity in a constant cycle of antagonistic coevolution. Long-term investigations carried out on samples from Chesapeake Bay revealed the larger role of bacteriophages in the microbial loop and carbon cycle [4]. Arguably, one of the most well characterized bacteriophage habitats is the marine [5]. This is due largely to the vast impact that marine cyanobacteria have in the oceans in their role as ecosystem engineers [6], and the subsequent impact that marine cyanophages affect upon their ecology. Despite this, the level of diversity in phage populations in freshwaters is likely to far exceed that found in marine environments, if only due to the known habitat heterogeneity found in freshwater systems [7]. Bacteriophages have a range of strategies for interacting with their hosts, including the potential to infect more than one bacterial genera, species, or strain [8,9], and variations in life-cycle after infection [10,11]. Investigating environments which are likely to support highly diverse phage populations are an important step towards examining host-phage interactions, therefore understanding more about how phages may drive important environmental processes.

Preliminary studies assessing phage diversity in the environment have involved the use of culture-based plaque assays. As with bacteria, these studies represented a negligible fraction of total phage abundance and heterogeneity present in the environment, and genomic techniques for examination are now becoming the standard approach. Standard methods for phage characterization are now, increasingly, environmental metagenomics and *de novo* whole genome sequencing (WGS) of isolated viruses [12]. While genomic approaches also have limitations and should be applied within the scope of strong experimental design, they have allowed for a glimpse into phage diversity and relative abundances for range of niches, e.g. human microbiota [13,14], soil [15] and aquatic samples: marine [16-19] and freshwater [20,21]. Regardless of the environmental niche investigated, the majority of sequence reads generated for virome projects show little homology to annotated function or taxonomy.

Testament to the transient evolutionary nature of bacteriophages, the largest issue faced by microbiologists studying genomic data derived from phage populations is the lack of a single marker gene, analogous to the 16S rRNA gene in prokaryotes. Furthermore, their simple biochemical composition belies the complex array of structural characteristics present in known viruses: double-stranded or single-stranded nucleic acid, RNA or DNA, circular, linear, or segmented (varying in size from just a few thousand nucleotides to hundreds of thousands of nucleotides in length), and a considerable range of possible morphological characteristics which may only be assessed effectively under high-powered microscopy by experienced specialists [22]. Individual studies of aquatic environments frequently narrow their scope to specific groups, e.g. RNA-based viruses [20, 23]) or DNA-based viruses [21, 24-26]. Studies of DNA-based viruses may include a rolling-amplification or multiple displacement amplification step, a technique which is known to introduce bias [27]. While efforts are now slowly being made to capture a more comprehensive estimation of the total viral diversity during sampling efforts [18], sample handling and processing remains complex and it is accepted that analysis by metagenomic assessment provides highly delineated information according to experimental design and the nature of the sampling effort.

In addition, while targeted viral studies alone provide insight into the genetic dynamism of aquatic environments, the paucity of existing information with which to compare such data requires, ultimately, direct comparison with similar evidence collected from bacterial populations on order to extend research beyond a catalogue of what may or not be present [28]. This is also important for beginning to assess how viruses of interest may interact with their host(s), and challenging or supporting current convention in understanding the ecology of viruses in the environment [29].

The authors of this study have previously investigated the temporal and spatial bacterial diversity present in the near waters of Chicago beaches, on the shores of Lake Michigan, one of the largest bodies of freshwater on the planet [30]. The sample collection and processing regime was designed such that a virally-sized fraction of water was also produced and assessed in order to establish preliminary information with regard to viral and bacterial diversity. This study presents an initial assessment of viral diversity and begins to couple the information with that of the bacterial cohort. Analysis has given indications of a rich source of viral genetic diversity in these samples, however, there is a clear lack of existing information with which to compare the vast majority of assigned operational taxonomic units (OTUs), as has been previously well-discussed. In light of this, this study also presents results obtained from open reading frame (ORF) predication, which supports the hypothesis that Lake Michigan represents a divergent reservoir for environmental viruses, which are likely present in highly complex relationships with the bacterial host population and mediated by a range of infection capacities and infective lifestyles.

BIG FINISH

**MATERIALS AND METHODS**

**Sample Collection, Summer 2013**

Two Chicago beaches, Montrose Beach (41.96675 / -87.63704) and 57th Street Beach (41.79072 / -87.57934), were selected as study sites given their relative equal proximity from city center and their varied qualities. No specific permits or permissions were required for the water samples collected from the Chicago Lake Michigan nearshore waters. The recreational swimming area of Montrose Beach, from which samples were taken, is abutted to the north by the Montrose Beach dog park and to the south by the Montrose Harbor Marina. The beach located at 57th Street is located on the south side of the city and is used solely for swimming. Sampling was conducted at both sites within the recreational swimming areas. Water was collected from the surface at a distance from the shore such that the water level was knee-height (approximately 0.5 m) deep. Four individual samples (4L each) were collected within a 5 m area from each site, the exception being on July 5th in which only three samples (4L each) were collected. This process was repeated every ten days from June 5th to August 14th, 2013 (16 samples in total).

Bacterial cells and viral particles were collected by filtering. The sample was first passed through sterile 0.45μm bottle-top cellulose acetate membrane filters (Corning Inc, Corning, NY) under vacuum to remove plant matter, sand, and debris, and eukaryotic cells. The process was repeated, using a 0.22μm polyethersulfone membrane filter (MO BIO Laboratories, Carlsbad, CA) to capture bacterial cells. Multiple filters were used per collection: the amount of water passed through each filter varied depending on the turbidity of the sample. To isolate viral particles a third filtration was conducted through a 0.10μm polypropylene filter (EMD Millipore Corp, Billerica, MA) using the Labscale™ tangential flow filtration (TFF) System (EMD Millipore Corp, Billerica, MA). Filtration was performed according to the manufacturer’s instructions.

**Viral DNA extraction and sequencing**

Viral DNA was extracted from all 16 samples using the MO BIO Laboratories PowerWater® DNA Isolation Kit (Carlsbad, CA). The protocol recommended by the manufacturer was followed with the exception of an additional heat treatment at 70°C for 10 minutes prior to initial vortexing. DNA isolated from each of the individual samples for a given collection date/location was pooled together. Concentrations were verified using the Qubit® Fluorometer (Life Technologies, Carlsbad, CA). DNA was stored at -20°C in ultrapure water until sequencing. To test against putative bacterial contaminants within the viral sample, each extraction assayed via PCR targeting the 16S rRNA gene with the 63f (5′-AGGCCTAACACATGCAAGTC-3′) and 1087r (5’-TCGTTGCGGGACTTACCCC-3′) primer pair. None of the extracted DNA samples produced bands while the positive control (*Escherichia coli* C DNA) did suggesting that the DNA isolated was viral (results not shown).

Nine of the 16 samples (Table 1) were selected for sequencing given their high DNA concentrations without any additional processing (e.g. rolling-amplification or multiple displacement amplification). Sequencing was performed by Genewiz (South Plainfield, NJ), using the Nextera XT kit for Illumina MiSeq library preparation, according to manufacturer’s guidelines. The nine samples were multiplexed for a single run of paired-end reads. Over 15 million raw reads were generated; the raw fastq files can be found via NCBI’s BioProject database (Accession: PRJNA248239).

**Paired-end read assembly and contig assembly**

Paired-end assembly and contig assembly was performed for each sample using MetaVelvet [32] with a hash length of k=29. The numbers of paired-end reads generated for each sample as well as the number of contigs generated for each sample are detailed in Table 1.

**Taxonomy classification and BLAST-based comparative viromics**

Taxonomic classification was conducted using the web server Metavir [33]. Due to the input limitations of Metavir, the longest 300,000 paired reads (Table 1) were selected for analysis. Metavir assesses virome composition using the GAAS tool [34]. BLAST analyses by Metavir were performed using the Refseq complete viral genomes protein sequences database (<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/viral/>); the release of the database used for comparison was that of Jan 18, 2014. This release contained 173,122 protein sequences. Our analysis of the sequences contained within the release includes ~65% of the proteins originating from phage genomes. In addition to taxonomic classification, Metavir computes rarefaction curves (Figure S1). Virome comparisons were also calculated for each sample to other publicly available freshwater viromes using Metavir. The Metavir method for virome comparisons, as described in Roux *et al*. [33], relies on tBLASTx comparisons using a subsample of the sequences pertaining to a virome.

**Comparative analysis of viromes based on contig assembly**

Each of the nine viromes was then assembled in a pair-wise fashion such that reads generated from one sample were assembled to the reads generated from another sample using the MetaVelvet tool [32]. This cross-sample assembly strategy can identify any similarities in species present within two samples.

**ORF generation**

The twenty longest contigs assembled by MetaVelvet for each sample were examined further. ORFs within each of these contigs were predicted using GeneMarkS (Besemer et al. 2001). Each predicted ORF was then BLASTed using the blastx algorithm via the NCBI web interface (results not shown). BLAST homologies having an e-value less than 10-3 were considered putative hits. BLASTs were not limited to only viral records but all records within the non-redundant protein sequences (nr) database. While some predicted ORFs produced significant hits, the majority of these hits were to coding regions annotated as “hypothetical proteins”. Furthermore, many of the predicted ORFs exhibited no sequence similarities to the nr database.

**Results AND DISCUSSION**

**Overview of Lake Michigan viromes**

The nearshore waters of two Chicago area beaches, Montrose Beach and 57th Street Beach, were sampled throughout the Summer of 2013. DNA isolated from viral particles purified from these nine samples was sequenced using standard methods, producing >15 million raw reads (see Methods). Taxonomic classification of the viral contigs was performed using the tool Metavir [33]. Metavir accepts sample submission to 300,000 contigs of lengths greater than 300 nucleotides. The submission cutoff defined our analysis to a subset of the contigs generated by sequencing. Therefore, for the purposes of the analyses detailed in this study, contigs were selected on the basis of length. Phages were found to comprise the majority of viruses present throughout the generated datasets. Rarefaction curves produced by Metavir revealed that the viral community from the 57th Street Beach on June 25th was contained within the contigs examined; in contrast, the curves of the remaining eight samples indicated significantly greater diversity (Supplemental Figure 1). Additional data collected from 57th Street Beach at the same time do not give any insight into why bacteriophage diversity may have been lower, and it is generally accepted that phage populations closely follow the composition and abundance of host bacterial populations in a given environment. However, Roux *et al*., (2012) described a similar level of discrepancy in rarefaction for two freshwater lake samples. In addition, as with earlier examples of similar sequencing efforts, the potential for sampling, sample preparation and sequencing bias is considerable [35]. There is a severe paucity of knowledge surrounding phage population dynamics and phage/host interactions in freshwaters, and therefore it is very difficult to interpret why one sample may have been so markedly different to others taken from the same environment, albeit at different times. However, the general heterogeneity that can be expected in freshwaters may suggest that such differences in diversity, on a temporal basis, are indeed possible, if not expected.

To date, the considerable majority of viral community assessments in the aquatic environment have been performed on marine samples. However, the genetic richness and diversity of bacteriophages found in freshwaters is likely to surpass that of their marine counterparts [36-37], and in congruence with this hypothesis, recent metagenomic assessments have demonstrated the presence of a wide range of ssRNA, dsRNA, ssDNA and dsDNA viruses in freshwater environments. For the purposes of preliminary assessment of viral diversity in the samples collected, this study was designed to assess the presence of DNA viruses. Due to the nature of sample preparation, i.e. no use of multiple displacement amplification (MDA - known to produce bias towards returning ssDNA viruses [27]), we estimated that the majority of BLAST hits returned by our sequencing effort would likely belong to dsDNA viruses. As with most studies examining viral metagenomics (SIV NEW REFS NOT AQUATIC), the vast majority of the sequences generated as part of this examination of the waters of Lake Michigan did not show any significant sequence similarity to cultured viral isolates: only 6.9% of the 2.6 million contigs submitted for analysis were found to produce a BLAST score of ≥50 to entries in the NCBI RefSeq viral protein database.

**Comparison Lake Michigan viromes**

The Lake Michigan viromes were first compared directly to the 20 viromes generated from freshwater environments (publicly available on Metavir). This was facilitated via the Metavir virome comparison functionality, and was based upon the BLAST results found for each of the viromes. As shown in Figure 1, the Lake Michigan samples are monophyletic. Moreover, the Lake Michigan viromes are distinct from the other 20 Metavir viromes. Although each of the Lake Michigan viromes consist of more sequences than several of the other viromes for which they are compared, normalization is performed by Metavir. Despite the inherent difficulties in directly comparing separate freshwater environments, due to the innate levels of heterogeneity each habitat is likely to exhibit, and despite the focus being on dsDNA bacteriophages, the dataset generated from this study appears to be novel, and clearly divergent from previous studies. There is no reason to suppose that the level of diversity seen in the viromes detailed as part of this study is not present throughout the groups of viruses not assessed by the sampling regime employed during this research. Indeed, other studies have previously found varied and distinctive populations of ssDNA viruses in freshwater habitats – the main difference from this one being the use of phi20 polymerase during MDA: known to bias for the amplification of this group, as previously discussed.

The BLAST-based comparisons captures only a small fraction of the genomic content within the viromes. The complete set of contigs generated for each of the nine Lake Michigan viromes was assembled in a pair-wise manner suing the assembler MetaVelvet [32] (see Methods). This cross-sample assembly strategy can identify any similarities in species present within two samples, representative of viral species and/or coding sequences present within both samples. In summary, the contigs produced by this process were far less numerous and shorter than those produced within the single sample assemblies (Figure 1B).

**Taxonomic classification of viromes**

The BLAST results for each of the nine samples included a variety of taxa (Figure 2), the most numerous belonging to the order Caudovirales, comprising the dsDNA families *Myoviridae*, *Siphoviridae* and *Podoviridae*. While the majority of hits were to dsDNA phages, similarities to ssDNA phages as well as eukaryotic-infecting viruses were also observed. In order to visualize the taxonomic composition of the BLAST hit results, the Krona tool [38], available as part of Metavir, was utilized. As shown in Figure 3 for the metagenome of the 57th Street Beach sample taken on June 5th, 96% of the proteins hits are to coding regions within dsDNA viruses (having no RNA stage), 60% of which were for *Caudovirales*. The Krona files for the remaining samples also illustrate this trend; all nine Krona files can be accessed and queried through the Metavir webserver (http://metavir-meb.univ-bpclermont.fr/). These data correlate strongly with existing literature for these groups of phages, previous metagenomic datasets for a range of viromic studies, and their representative presence in the BLAST database (REFs).

While the BLAST hits related to proteins belonging to a variety of viral species, 1182 species produced at least one hit to contigs belonging to all nine samples. The majority (949, 80%) of these hits are to specific phages. As the Krona pie charts show, some of the hits to individual species are significantly greater than others. This does not, however, necessarily mean that specific phages are present in the sample in any considerable abundance. For example, all nine samples showed numerous hits to the bacteriophage Planktotrixphage PaV-LD (NC\_016564). Further investigation of these hits, however, showed that nearly all of these hits (over 7000 in the nine samples; Figure 4A) are to the gene PaVLD\_ORF033R; this structural protein is annotated as an ABC transporter protein [39]. This putative abundance is extremely unlikely to represent a single ‘species’ of phage only. The hit, therefore, is more likely due to the high presence of ABC transporter proteins (ubiquitous throughout all extant prokaryotic and eukaryotic phyla alike), in the sample generally. However, the abundant presence of a potentially non species-specific protein in a dataset such as this is not a useless artifact. Rodriguez *et al*., (2010) found a comparatively low abundance of ABC transporter proteins in a freshwater microbiome. This may relate to differences in sampling methods, or general community characteristics. Considering the Lake Michigan virome demonstrates a discrete phylogenetic cluster (Figure 1) when compared to other datasets available on Metavir, we can expect such disparity in the presence of key functional proteins. Such characteristics may be used to compare dynamic samples from the same location, or, as in this case, from those which may be entirely different.

The percentages represented in the Krona pie charts do not represent the coverage of hits to an individual species: this information provides greater insight into the likelihood of the species present, as opposed to the gene(s) present. Therefore, BLAST query results were mined considering both the number of hits to the genome as well as the percentage of annotated genes in the genome with one hit or more. All nine samples generated numerous hits to phages infectious to *Burkholderia*. For example, fifty percent of the 72 annotated protein-coding genes within the 47,399 bp genome of the *Burkholderia* phage BcepB1A (NC\_005886) [40] were found within the Lake Michigan samples. As shown in Figure 4B, the majority of the hits are from three samples – the samples taken on June 5th at both beaches and the Montrose Beach sample from June 15th. Samples taken later in the summer had few if any hits to this genome; the sample from 57th Street Beach on June 25th had no hits to this particular *Burkholderia* phage, although numerous hits to other phage infecting the same host - *Burkholderia cepacia*. High coverage for phage species known to possess larger genomes was also observed, for example, the 252,401 bp genome of *Prochlorococcus* phage P-SSM2 (NC\_006883) shown in Figure 4C. While *Prochlorococcus* is a marine cyanobacterium, it has been determined previously that P-SSM2, a myovirus, is likely to contain a number of core genes, both belonging to T4-like groups and those which are cyanobacterial in origin [41]. This lends further support, to some extent, to a ‘gene hunting’, as opposed to a ‘phage hunting’ approach in viral metagenomic datasets – *Prochlorococcus* is unlikely to be present in Lake Michigan, as a marine microorganism, but a vast catalogue of freshwater cyanobacteria will be indigenous and present in association with freshwater cyanophages. Both are likely to possess the genes identified in our Lake Michigan samples.

Hits to eukaryote-infecting viruses were often few, 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 mimivirus, moumouvirus, megaviruses and pandoraviruses. For instance, over 16% of the annotated genes in *Acanthamoeba polyphaga mimivirus* (NC\_014649) were found amongst the nine metagenomes produced here. The Lake Michigan viromes had numerous hits to the Cation Channel-forming Heat-Shock Protein-70 (YP\_003986897) (Figure 4D).

We can therefore validate our approach (i.e. assessment of key differences in important groups of genes) in several ways: by demonstrating overrepresentation of a group (the *Planktothrix* phage), demonstration of variability in the presence of a specific phage group across samples (the *Burkholderia* phages) and by showing the presence of ubiquitous viruses throughout the sampling effort (the giant viruses).

While hits to the ABC transporter coding region (YP\_004957306) were previously specifically observed for the *Planktothrix* phage PaV-LD genome (Figure 4A), 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 contig and the RefSeq database was localized to the ABCC\_MRP\_Like domain (CDD ID: cd03228). Expanding our analysis to all hits, not just those to phage species, there were a significant number analogous to coding regions annotated as an ABC transporter within the genomes of amoeba-infecting viruses. As the Metavir analysis is limited to the RefSeq annotated viral genome collection, we were interested 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 (nr) protein sequences database protein 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, including those from the Organic Lake study [42]. Likewise the BLASTp search to the env\_nr database identified statistically significant hits (E-value = 8e-37) to sequences generated from global ocean sampling studies. These analyses reconfirm that in a dataset as complex as an aquatic viromes, where the vast majority of BLAST hits are non-viable, the presence of specific genes is what may give insight into community structure, as opposed to the species assigned to that gene specifically.

**Detection of viral-associated proteins**

Comparing sequence composition of contigs to existing annotated metagenomic information can also give us insight into the functional capacity of viruses: and begin to suggest interactions with host bacteria within the complex community of the nearshore waters of Lake Michigan. The BLAST hits to phage species produced by the Metavir analyses for each of the nine viromes were aggregated by function (based upon their annotated protein product). The number of BLAST hits per annotated function varied considerably (91±635). Those functions with the greatest number of hits were further investigated. As expected, and again, as is typical of similar datasets, the majority of generated ORFs were assigned to hypothetical proteins (Table 2). For those ORFs which returned more descriptive hits (including DNA helicases, DNA integrases, DNA polymerases, portal proteins, structural proteins and the large subunit of terminases), all were observed to belong to those proteins most prevalent in single annotated phage genomes [43]. Interestingly, integrases are highly ubiquitous, and can be shared by phages, whereas previous bioinformatic analysis [43], found that many of these functions are phage-specific, including the large subunit of terminases, portal proteins, and structural proteins. Other phage-specific ORFs frequently present within annotated viral genomes, for example the cro/cI repressor (important in lytic and lysogenic development [44] and homing HNH endonucleases (important in lateral transfer [45]) were also identified in the Metavir hits, albeit far fewer in number (1315 and 3855, respectively). This suggests a strong representation throughout our dataset, and others (REFs), of phage/prophage populations, as opposed to genes shared by hosts/phage present in bacterial DNA.

**Virus-Host dynamics**

The dependence of viral species on their host(s) for propagation mandates that environmental phages and their hosts should be closely associated in the wild. The use of metagenomic techniques to examine community diversity can be a powerful technique, however, few studies have begun to link the information generated from viromic datasets to phage/host interactions: a crucially important aspect of the impact that phages have on the environment. The often cyclical battle between virus and host can reduce the population of either below the threshold able to be detected via direct whole genome sequencing. Moreover, lysogenic phage present within the environment cannot be directly assessed via the isolation strategy employed here. Through direct identification of the bacterial taxa present within a sample it is possible to ascertain if viral taxa predicted by, for instance, BLAST are *likely* to be present within a sample. In previous work [30], the same waters investigated here for viral presence were assessed using a metagenetic survey of 16S rRNA (Figure 5 Left Panel). For each of the viral taxa identified using the BLAST classification tool in Metavir, the host species was identified (Table S1). Putative relative abundances of host species were then calculated based upon the number of hits to a particular viral strain (Figure 5 Right Panel).

The nearshore waters of Lake Michigan have abundant populations of Proteobacteria and viruses which infect Proteobacteria (Figure 5). While 16S rRNA analysis reveals significant populations of Bacteroidetes, particularly for the collection from Montrose Beach on June 25, the number of hits to viruses infecting Bacteroidetes spp. is relatively few (1.70-4.75% of BLAST hits/ sample). The viral hosts included within the “Other” category of the graph include both Firmicutes and Cyanobacteria. While 16S rRNA survey data detected the presence of Firmicutes, it did not identify any putative members of the Cyanobacteria. This is a confounding result, considering the presence of cyanobacteria in most freshwater environs is considerable [46]. However, the presence of hits pertaining to cyanobacteria according to the presence of *cyanophage associated sequences* is considerable. This difference may be due to a number of factors – for example, in a mixed bacterial sample, cyanobacterial nucleic acids are difficult to efficiently extract and therefore their presence may be masked in a sample such as this. This is also true of a number of other groups of bacteria. The bacterial phyla abundance predictions, based upon the BLAST viral hits, however, did show strong correlation to the annotated hosts of viral species within the RefSeq database, despite the RefSeq collection having far fewer closed and annotated genomes for Bacteroidetes-infecting phage than Proteobacteria-infecting species.

**Conclusions**

BIG FINISH

**Acknowledgments**

This research was funded by the NSF (1149387) (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 phage-host data.

**REFERENCES**

**\*add Malki et al 2014**

1. Claverie
2. La Scola B, Desnues C, Pagnier I, Robert C, Barrassi L, et al. (2008) The virophage as a unique parasite of the giant mimivirus. Nature 455: 100–104.
3. Rice G, Tang L, Stedman K, Roberto F, Spuhler J, et al. (2004) The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. Proc Natl Acad Sci USA 101: 7716–20.
4. Winget D, Helton R, Williamson K, Bench S, Williamson S, et al. (2011) Repeating patterns of virioplankton production within an estuarine ecosystem. Proceedings of the National Academy of Sciences of the United States of America
5. Suttle CA. (2005) Viruses in the sea. Nature 437: 356–61.
6. Berman-Frank I, Lundgren P, Falkowski P. (2003) Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. Res Microbiol 154: 157–64
7. Hyman P, Abedon ST. (2009) Practical methods for determining phage growth parameters. In: Clokie MRJ & Kropinski AM, editors. Bacteriophages: Methods and Protocols 1: Isolation, Characterization and Interactions. New York: Humana Press. pp. 175-202.
8. Ellen C. Jensen, Holly S. Schrader, Brenda Rieland, Thomas L. Thompson, Kit W. Lee, et al. (1998) Prevalence of broad host range lytic bacteriophages of *Sphaerotilus natans*, *Escherichia coli* and *Pseudomonas aeruginosa*. Appl. Environ. Microbiol. 64: 575-580.
9. Watkins S, Smith J, Hayes P, Watts J (2014) Characterisation of Host Growth after Infection with a Broad-Range Freshwater Cyanopodophage. PLoS ONE 9: e87339.
10. McDaniel L, Breitbart M, Mobberley J, Long A, Haynes M, et al. (2008) Metagenomic analysis of lysogeny in Tampa Bay: Implications for prophage gene expression. PLoS One 3: 1-9.
11. Holmfeldt K, Solonenko N, Shah M, Corrier K, Riemann L, et al. (2013). Twelve previously unknown phage genera are ubiquitous in global oceans. Proc Natl Acad Sci 110: 12798-12803
12. Reyes A, Semenkovich N, Whiteson K, Rohwer F, Gordon J (2012) Going viral: next-generation sequencing applied to phage populations in the human gut. Nature reviews Microbiology 10: 607–617.
13. Minot S, Sinha R, Chen J, Li H, Keilbaugh S, et al. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome research 21: 1616–1625
14. Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, et al. (2007) Metagenomic and Small-Subunit rRNA Analyses Reveal the Genetic Diversity of Bacteria, Archaea, Fungi, and Viruses in Soil. Applied and Environmental Microbiology 73: 70597066.
15. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: 2121-2131.
16. Williamson S, Allen L, Lorenzi H, Fadrosh D, Brami D, et al. (2012) Metagenomic exploration of viruses throughout the Indian Ocean. PloS one 7: e42047
17. Hurwitz B, 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.
18. Mizuno C, Rodriguez-Valera F, Kimes N, Ghai R (2013) Expanding the Marine Virosphere Using Metagenomics. PLoS Genetics 9: e1003987.
19. Djikeng A, Kuzmickas R, Anderson N, Spiro D (2009) Metagenomic analysis of RNA viruses in a fresh water lake. PloS one 4: e7264.
20. Roux S, Enault F, Robin A, Ravet V, Personnic S, et al. (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics.
21. Ackermann H-W (2013) Sad State of Phage Electron Microscopy. Please Shoot the Messenger. Microorganisms 2: 110
22. Bolduc B, Shaughnessy D, Wolf Y, Koonin E, Roberto F, et al. (2012) Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs. Journal of virology 86: 5562–5573.
23. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452: 340–343.
24. Lopez-Bueno A, Tamames J, Velasquez D, Moya A, Quesada A, et al. (2009) High diversity of the viral community from an Antarctic lake. Science 326: 858–861.
25. Roux S, Enault F, Robin A, Ravet V, Personnic S, et al. (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics.
26. Williamson S, Allen L, Lorenzi H, Fadrosh D, Brami D, et al. (2012) Metagenomic exploration of viruses throughout the Indian Ocean. PloS one 7: e42047
27. Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, et al. (2012) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. The ISME Journal 7: 359369.
28. Hurwitz B, 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.
29. Mizuno C, Rodriguez-Valera F, Kimes N, Ghai R (2013) Expanding the Marine Virosphere Using Metagenomics. PLoS Genetics 9: e1003987.
30. Kim K-H, 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
31. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, et al. (2010) Viral and microbial community dynamics in four aquatic environments. The ISME journal 4: 739–751.

**Table 1.** Summary statistics of sequencing.

|  |  |  |  |
| --- | --- | --- | --- |
| **Date** | **Site** | **Number of Paired-End Reads** | **Number of Assembled Contigs** |
| Jun 5, 2013 | Montrose Beach | 1,123,802 | 565,244 |
| Jun 5, 2013 | 57th Street Beach | 2,202,471 | 722,253 |
| Jun 15, 2013 | Montrose Beach | 1,380,226 | 808,040 |
| Jun 25, 2013 | Montrose Beach | 1,547,157 | 1,049,321 |
| Jun 25, 2013 | 57th Street Beach | 1,626,554 | 81,777 |
| Jul 5, 2013 | Montrose Beach | 1,854,527 | 1,199,969 |
| Jul 15, 2013 | 57th Street Beach | 2,166,575 | 948,441 |
| Jul 25, 2013 | Montrose Beach | 2,351,827 | 1,428,015 |
| Aug 14, 2013 | 57th Street Beach | 1,415,911 | 648,250 |

**Table 2.** Annotations of proteins most frequently observed within the Lake Michigan Virome contigs.

|  |  |
| --- | --- |
| **Annotated Protein Function** | **# Hits to Lake Michigan Viromes (% of all hits)** |
| Hypothetical/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%) |

Figure Legends

**Figure 1.** Comparison of Lake Michigan viromes to **(A)** other freshwater viromes based upon sequence similarities. **(B)** The tree represents the expansion of the Lake Michigan Viromes node in panel A. The heatmap illustrates the similarities/differences between the nine viromes quantified by cross-sample assembly.

**Figure 2.** Taxonomic classifications identified via BLAST searches between the June 5, 2013 57th Street Beach virome contigs and RefSeq viral proteins.

**Figure 3.** Coverage of Lake Michigan virome contigs across the genomic sequence of (A)

**Figure 4.** Comparison of Lake Michigan viromes to other freshwater viromes based upon sequence similarities

**Figure 5.** Relative abundances of bacteria phyla as assessed via (left) 16S rRNA sequencing and (right) annotated hosts for BLAST-based viral taxa predictions.