Freshwater Metaviromics and Bacteriophages: A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges

Katherine Bruder  
*Loyola University Chicago*

Kema Malki  
*Loyola University Chicago*

Alexandria Cooper  
*Loyola University Chicago*

Emily Sible  
*Loyola University Chicago*

Follow this and additional works at: [https://ecommons.luc.edu/bioinformatics_facpub](https://ecommons.luc.edu/bioinformatics_facpub)

See next page for additional authors

Recommended Citation  
Bruder, Katherine; Malki, Kema; Cooper, Alexandria; Sible, Emily; Shapiro, Jason W.; Watkins, Siobhan C.; and Putonti, Catherine. Freshwater Metaviromics and Bacteriophages: A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges. Evolutionary Bioinformatics, Suppl 1, : 25-33, 2016. Retrieved from Loyola eCommons, Bioinformatics Faculty Publications, [http://dx.doi.org/10.4137/EBO.S38549](http://dx.doi.org/10.4137/EBO.S38549)

This Article is brought to you for free and open access by the Faculty Publications at Loyola eCommons. It has been accepted for inclusion in Bioinformatics Faculty Publications by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.

This work is licensed under a Creative Commons Attribution-No Derivative Works 3.0 License. © The authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC3.0 License.
Authors
Katherine Bruder, Kema Malki, Alexandria Cooper, Emily Sible, Jason W. Shapiro, Siobhan C. Watkins, and Catherine Putonti

This article is available at Loyola eCommons: https://ecommons.luc.edu/bioinformatics_facpub/21
Freshwater Metaviromics and Bacteriophages: A Current Assessment of the State of the Art in Relation to Bioinformatic Challenges

Katherine Bruder¹, Kema Malki¹, Alexandria Cooper¹, Emily Sible¹, Jason W. Shapiro¹,², Siobhan C. Watkins¹ and Catherine Putonti¹–³

¹Department of Biology, Loyola University Chicago, Chicago, IL, USA. ²Bioinformatics Program, Loyola University Chicago, Chicago, IL, USA. ³Department of Computer Science, Loyola University Chicago, Chicago, IL, USA.

Supplementary Issue: Bioinformatics Methods and Applications for Big Metagenomics Data

ABSTRACT: Advances in bioinformatics and sequencing technologies have allowed for the analysis of complex microbial communities at unprecedented rate. While much focus is often placed on the cellular members of these communities, viruses play a pivotal role, particularly bacteria-infecting viruses (bacteriophages); phages mediate global biogeochemical processes and drive microbial evolution through bacterial grazing and horizontal gene transfer. Despite their importance and ubiquity in nature, very little is known about the diversity and structure of viral communities. Though the need for culture-based methods for viral identification has been somewhat circumvented through metagenomic techniques, the analysis of metaviromic data is marred with many unique issues. In this review, we examine the current bioinformatic approaches for metavirome analyses and the inherent challenges facing the field as illustrated by the ongoing efforts in the exploration of freshwater phage populations.

KEYWORDS: metaviromics, environmental metagenomics, bacteriophages, freshwater

Introduction

Advances in sequencing technologies over the past two decades have led to many studies examining the complex microbial communities that support life on Earth. From extreme environments¹ to buildings,² surveys of microbial communities have identified many novel taxa of bacteria and archaea.³,⁴ In parallel, exploration of the viral fraction of microbial communities has also consistently uncovered novel genetic content.⁵,⁶ With an estimated 10³⁰⁰ viruses present in the ocean alone,⁷ and considering our current surfeit of knowledge,⁸ viruses represent a vast untapped reservoir of genetic diversity.⁹,¹⁰ Viruses that infect bacteria (bacteriophages) are some of the most abundant biological entities on the planet: they play a critical role in shaping microbial community structure and metabolism through mediation of mortality and genetic mobility.¹¹ Therefore, phages have a global impact on biogeochemical nutrient cycling.¹²

Assessing the diversity of viral species is not as straightforward as it is for prokaryotes. While bacteria and archaea can be classified via the 16S rRNA gene, there is no single conserved gene among all viral species. Thus, whole-genome sequencing (WGS), rather than targeted gene sequencing, must be employed to begin to assess the heterogeneity of viral taxa present within a sample (the metavirome). The WGS approach has also been employed in numerous studies of bacterial and archaeal communities¹³–¹⁵; these studies have a crucial advantage – a well-populated repository of characterized gene sequences. Despite the critical role phages play in the natural environment as well as within the human microbiota,²,¹⁶–²¹ data repositories lack sufficient quantity and diversity of phage gene and genome sequences. This is due to a general dearth of phage research in comparison to that of all other organisms, as well as the challenges associated with working with phages in the laboratory. Predominantly, the isolation of phages is limited in the same way as that of bacteria – only a fraction can be successfully isolated from the environment and maintained in laboratory settings. In addition, direct isolation and the use of metagenomics-based analyses are heavily biased in favor of examining the communities of phages that are actively destroying bacterial cells. These lytic phages are present as unattached...
viral particles in the environment. However, a large cohort of phages integrate with host genomes in a state known as lysogeny, forming prophages. Prophages are often overlooked during metaviromic examinations, along with all other bacterial data. Furthermore, many prophages are mislabeled in data repositories as being innately bacterial.\textsuperscript{22,23}

Interest in phage ecology and community dynamics within natural and man-made environments continues to accelerate.\textsuperscript{24} Herein, we review the current bioinformatic approach for metavirome analyses with specific focus on the study of complex phage communities. While phage communities have been investigated in a variety of ecosystems, we focus on phage communities in freshwaters. Freshwater represents a potentially highly dynamic and variable community of phages and is likely to support a novel cohort of genetic and phenotypic diversity. In comparison to the marine environment, freshwater phages have been understudied: in light of the limitations already discussed, this paucity of data contributes to our currently restricted insight into these important communities. Such hindrances reflect the general technicalities that are associated with the analysis of metagenomic data, which are, inarguably, very valuable to our broader knowledge of viral and microbial communities in the environment. Herein, we begin to examine the practical aspects of bioinformatics-based analyses of metavirome data, particularly those from freshwater, and how solutions to associated problems may inform the broader field.

**Bioinformatic Analysis of Environmental Metavirome Datasets**

**Current methodology.** It is common practice to follow the well-established protocols developed for the investigation of microbial communities.\textsuperscript{25} Prior to beginning analyses, raw viral sequencing reads should be inspected to remove sequencing artifacts (eg, primers and/or adaptors). While contemporary sequencing platforms typically have very low overall error rates, errors and biases do occur\textsuperscript{26} and can be identified and removed alongside low-quality bases via a number of tools [eg, FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)]. Given the importance of high-quality data, several open-source software tools have been developed to facilitate quality control (eg, HTQC\textsuperscript{27}; SUGAR\textsuperscript{28}; khmer\textsuperscript{29}).

While there are many methods for processing WGS datasets, extant metaviromic studies – particularly those focused on complex phage communities – typically follow a protocol similar to that shown in Figure 1. The reads produced, irrespective of the sequencing technology used, are first assembled into contigs using tools such as Velvet\textsuperscript{30} and SPAdes.\textsuperscript{31} While Figure 1 lists some of the assemblers frequently used within freshwater metavirome studies, this is in no way an exhaustive representation of the tools available.\textsuperscript{25,32,33} Postassembly, the contigs can either be directly compared with viral data collections or open reading frame (ORF) prediction can be performed. In the latter case, several different software tools are publicly available, including those listed in Figure 1. Contigs are classified through heuristic homology comparisons (typically BLAST\textsuperscript{34} or in some cases BLAT\textsuperscript{35}) to available viral sequence collections, such as GenBank,\textsuperscript{36} RefSeq,\textsuperscript{37} and SEED.\textsuperscript{38} Downstream analyses and comparisons between viromes often rely on these homology-based results.

There are several cloud-based or remote services, which encapsulate the process outlined in Figure 1. While both MG-RAST\textsuperscript{39} and MEGAN\textsuperscript{40} were developed for analyses of bacterial community sequencing efforts, they can be applied to metavirome data analyses. Two tools, MetaVir\textsuperscript{41} and VIROME,\textsuperscript{42} have been designed specifically for the analysis of metaviromic data sets. Sequences are classified by MetaVir through the top blastx comparisons to the RefSeq Virus database.\textsuperscript{41} VIROME classifies viral sequences via the top blastp comparisons to the UniRef100 peptide database.\textsuperscript{42}

**Limitations of mainstream methodologies.** The ability to classify a sequence originating from a phage, whether attempting to identify the taxa present or the putative functionality of a coding region, is dependent upon the availability of representative sequences within the data repository used. As of January 2016, EBI’s Phage Genome collection (www.ebi.ac.uk/genomes/phage.html) includes only 2,010 organisms, similar to that reported in the RefSeq collection through NCBI (2,018 total). NCBI’s Nucleotide database (www.ncbi.nlm.nih.gov/nuccore) has just over 12,000 phage sequences (complete and partial), nearly three orders of

---

**Figure 1.** General protocol for metavirome data analyses. Commonly used resources are listed for individual steps.
magnitude less than the size of bacterial sequence collections. Thus, analyses of metaviromic datasets must acknowledge the dependency upon available sequence databases and the limited number of characterized species. The small fraction of phage diversity that has been sequenced also has inherent bias. dsDNA phage genomes outnumber those of RNA genomes, with Mycobacteriophages being the most expansively investigated, largely due to the work of PhageHunters (Fig. 2). Likewise, phages that infect the commonly encountered type strain hosts (ie, *Pseudomonas* and members of Enterobacteriaceae) are also comparatively overrepresented (Fig. 2). This imbalance within the databases presents a challenge when assigning importance to previously unclassified viral sequences from uncultured viral samples.

Furthermore, many sequence annotations are incorrect. Incomplete or mislabeled viral sequences in public databases are common. For instance, a blastx search of nucleic acid from purified virus-like particles against GenBank was shown to produce matches to nonviral sources. However, when these same sequences were compared with the ACLAME database (a collection of classified mobile genetic elements including phages, plasmids, and transposons), the majority of reads could be reclassified as plasmid or phage. Two tools have been developed to assist in identifying sequences of viral origin within microbial genomes: VirSorter and Phage-Phisher. Both can identify lysogenic viral sequences as well as prophage sequences. Prophages within bacterial genomic sequences can lead to erroneous classification of metaviromic sequences as being bacterial in origin, and thus necessitates further, often manual, investigation of such hits. Limiting comparisons to only annotated viral sequences circumvents this challenge; however, this comes at a cost of ignoring phage sequences that have only been identified within bacterial genomes (ie, their species of origin has yet to be discovered and characterized).

In addition to practical aspects of computational analyses, bias is also introduced during isolation and sample preparation. Filtering, a standard aspect of water sample preparation, potentially excludes large dsDNA viruses. Additionally, chloroform treatment, CsCl gradients, and random PCR amplification favor chloroform-tolerant viruses, tailed phages, and abundant taxa, respectively. Amplification of whole DNA via techniques such as multiple displacement amplification is known to favor single-stranded circular DNA in the amplification process, which may result in the overrepresentation of these viruses in some metaviromic studies. Sample storage times and temperatures may also exclude some environmental viruses as their decay rates vary.

**Exploring metaviromes – case study: freshwater samples.** The initial discovery that phages were highly abundant in aquatic samples paved the way for the eventual determination of the pivotal impact that phages have on global bacterial-mediated processes. Phage metagenomics essentially began in the marine environment, and marine datasets continue to be one of the most comprehensively examined from a biological and bioinformatic perspective (eg, The Pacific Ocean Virome). There are comparatively few datasets collected from freshwaters, despite their importance as sources of drinking water, recreation, and commerce. At the microbial scale, water chemistry and hydrological factors can contribute to a dynamic environment, which is likely to be reflected in the indigenous phage populations.

To date, there have been 12 freshwater metavirome datasets, which focus on phage populations within their samples and for which details regarding their methods and results of analyses are published (Table 1 and Fig. 3). Many of these studies have been performed using 454 sequencing, as presented in Table 1. In addition to the studies listed in Table 1, there have been several other metavirome datasets generated. A search through, eg, NCBI’s Sequence Read Archive (SRA: http://www.ncbi.nlm.nih.gov/sra/) or the iMicrobe data commons (http://data.imicrobe.us/), will reveal additional sequencing datasets from various other freshwater environments. Given our motivation here is on the analyses of metaviromic datasets, we will thus focus on those studies accompanied by publications.

As presented in Table 1, the majority of the sequences generated by metaviromic studies exhibit no discernible similarity to the sequences in current data repositories; in fact, at most 29.5% of the sequences generated exhibit homology to a known viral sequence. This is universal, regardless of the freshwater ecosystem under investigation. Freshwater metavirome surveys have largely been focused on DNA viruses, which include nucleic acids from both phages and eukaryotic viruses. However, phages constitute the majority of the identifiable sequences within these datasets. Nevertheless, from the small fraction which is identifiable, varying conclusions have been drawn.

![Figure 2. Composition of the annotated host(155,103),(845,925)
<table>
<thead>
<tr>
<th>SAMPLE SITE</th>
<th>DATA LOCATION</th>
<th>SEQUENCING TECHNOLOGY</th>
<th>DATASET SIZE (Mbp)</th>
<th>ANALYSIS METHOD</th>
<th>% OF SEQUENCES ABLE TO BE IDENTIFIED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lake Limnopolar (RNA)(^{53})</td>
<td>SRA: SRP044919</td>
<td>454</td>
<td>162.9</td>
<td>BLAST(^{34})</td>
<td>5.3–15.3</td>
</tr>
<tr>
<td>Lake Limnopolar (DNA)(^{45})</td>
<td>SRA: SRP000593</td>
<td>454</td>
<td>24.6</td>
<td>BLAST(^{34}) and MEGAN(^{40})</td>
<td>12.4</td>
</tr>
<tr>
<td>Lake Bourget and Lake Pavin (DNA)(^{60})</td>
<td>SRA: ERP000339</td>
<td>454</td>
<td>706.2</td>
<td>MetaVir(^{41})</td>
<td>14.3–26.4</td>
</tr>
<tr>
<td>Lake Needwood (RNA)(^{24})</td>
<td>NCB: ADVU0000001.1, ADVT000001.0</td>
<td>454</td>
<td>6.7</td>
<td>BLAST(^{34}) and MEGAN(^{40})</td>
<td>34 (included hits to viral, bacterial, archaeal, and eukaryotic sequences)</td>
</tr>
<tr>
<td>Lake Michigan (DNA)(^{48})</td>
<td>SRA: SRP042189</td>
<td>Illumina</td>
<td>23.7</td>
<td>2013 Samples: MetaVir(^{11}) 2014 Samples: BLAST(^{34}) (unpublished)</td>
<td>2013 Samples: 6.9</td>
</tr>
<tr>
<td>Florida potable water and reclaimed water (DNA and RNA)(^{46})</td>
<td>SRA: SRP000673</td>
<td>454</td>
<td>371.6</td>
<td>BLAST(^{34})</td>
<td>Potable DNA: 56 Reclaimed DNA: 30 Reclaimed RNA: 43 (included hits to viral, bacterial, archaeal, and eukaryotic sequences)</td>
</tr>
<tr>
<td>Feitsui Reservoir (DNA)(^{58})</td>
<td>SRA009395</td>
<td>454</td>
<td>119.9</td>
<td>BLAST(^{34})</td>
<td>31.7</td>
</tr>
<tr>
<td>4 Kent Sea Tilapia Ponds (DNA)(^{9})</td>
<td>SRA: SRP000139, SRP000138, SRP000163</td>
<td>454</td>
<td>73.1</td>
<td>BLAST(^{34})</td>
<td>2.9</td>
</tr>
<tr>
<td>Ponds of the Mauritanian Sahara (DNA)(^{57})</td>
<td>MG-RAST: 444603.3, 4445718.3, 4445716.3, 4445715.3</td>
<td>454</td>
<td>82.8</td>
<td>MG-RAST(^{39})</td>
<td>16.79–29.5</td>
</tr>
<tr>
<td>Lake Ontario and Lake Erie (DNA)(^{31})</td>
<td>SRA: SRP060006</td>
<td>Illumina</td>
<td>4986</td>
<td>MetaVir(^{11}) and MG-RAST(^{31})</td>
<td>−25</td>
</tr>
<tr>
<td>Artic fresh water viromes (DNA)(^{74})</td>
<td>SRA: ERS396648</td>
<td>Illumina454</td>
<td>−7300</td>
<td>BLAST(^{34}) and MEGAN(^{40})</td>
<td>9.8</td>
</tr>
<tr>
<td>Lake Matoaka (DNA)(^{34})</td>
<td>MetaVir(^{11})</td>
<td>454</td>
<td>48.4</td>
<td>MG-RAST(^{39}) and MetaVir(^{11})</td>
<td>23.9–25.7</td>
</tr>
<tr>
<td>Lake Lough Neagh (DNA)(^{79})</td>
<td>SRA: SRP062094</td>
<td>Illumina</td>
<td>−1200</td>
<td>MetaVir(^{11}) and MG-RAST(^{39})</td>
<td>14.6 (MetaVir)−15% (MG-RAST)</td>
</tr>
</tbody>
</table>

**Note:** Locations of sites sampled are shown in Figure 3.
Nutrient levels affect phage community structure. Previous work suggests that the diversity of DNA phages in freshwater environments is subject to change based on various environmental factors including temperature and available nutrient levels. Freshwaters of lower nutrient availability tend to have less viral species richness in comparison to their higher trophic counterparts, e.g., the mesotrophic Lake Bourget and oligotrophic Lake Pavin. This corresponds to greater taxonomic diversity of bacterial hosts within environments with high nutrient availability. However, this contradicts a study performed on Antarctic samples from the oligotrophic Lake Limnopolar, which demonstrates high species richness. Given the fact that the databases, from which species identification is made, are themselves not representative of phage diversity, claims of changes in diversity are inherently biased. Furthermore, as the vast majority of the virome is unidentifiable, quantifying the true diversity present is not possible.

Variation due to seasonality and weather. Correlations between season and species diversity have also been observed. Furthermore, shifts in the taxa have been associated with environmental stressors, including decreased rainfall. In a study examining Saharan gueltas (ponds), the authors postulated that the extreme conditions of this site favor lysogenic phages, supported by their observation of Microbacterium phage Min1. Given phage dependence upon the presence and susceptibility of their bacterial host(s) and the seasonality previously observed within bacterial species in freshwaters, fluctuations in viral population structure are expected. These results rely on the small fraction of identifiable sequences. Exploring the unknown constituent is far more difficult. Cross-sample assemblies – assembling contigs from one sample with another sample – can provide some insight into the similarity/dissimilarity between sequences, regardless of their homology to databases.

Spatial and temporal variation. Examination of freshwater bodies from the Arctic and Antarctic revealed similar taxa within their waters, despite their geographic distance. Similarly, when the metaviromes of Lake Ontario and Lake Erie were examined concurrently, little to no significant difference in species diversity was observed. However, temporal variation was observed in both lakes when taxonomic composition was examined between samples taken a year apart. For example, although dominant in the collections of 2012, in 2013 Microviridae populations decreased in both Lake Ontario and Lake Erie samples. While these two studies compare somewhat similar habitats, comparison between diverse environments (such as those presented in Table 1) highlights the variability observed. Nevertheless, the same caveats discussed earlier are just as true here.

Figure 3. Locations of the freshwater metaviromes in Table 1.
Anthropogenic effects. The effects of human-related impact on the microbial environment have recently been attracting more attention. In a study of the Virginian Lake Matoaka, viral species richness and diversity was found to be negatively correlated with the level of human activity at the sample site, with the highest levels of diversity and species richness found at the main body of the lake, suggesting the area least affected by human activity. Similar results were found in Saharan gueltas; the guelta most influenced by human beings exhibited the lowest amount of viral diversity and more heterotrophic microorganisms and human pathogens. When potable and reclaimed water were compared in terms of viral abundance, it was found that the latter contained 1,000-fold more viruses than the former; reintroducing reclaimed water within the environment can thus potentially affect native microbial species.

Caveats of Applying Existing Methods of Analysis to Environmental Metaviromic Datasets

As the discussed studies of freshwater phage populations exemplify, the dearth of characterized phage genome sequences limits the ability to classify the majority of metaviromic sequences. The genomes of phages are extremely plastic and are able to shuttle genes between organisms; looking at the number of BLAST hits to a single genome may erroneously indicate the species’ presence. For instance, several thousand hits to the Planktothrix phage PaV-LD genome suggested the presence of this phage within the Lake Michigan nearshore waters. However, further investigation of these BLAST results revealed that these hits were to a single gene, thus suggesting that the gene rather than the species was present within the sample. Thus, BLAST is an effective means of assessing the presence and absence of genes of interest; similarly, metaviromic sequences could be mapped to user-selected gene sequences to detect and qualify the prevalence of genes of interest. Nevertheless, in the absence of a conserved genetic marker, ascertaining the presence of particular species within metaviromic sequences is fraught with challenges that have yet to be addressed.

As metaviromic surveys of environments continue, it is imperative – regardless of the environment sampled – that any conclusions drawn with regard to incidence of particular viral species be informed, considering not only the number of hits to a particular reference genome but also its coverage (the percentage of a query sequence that is aligned with a database sequence). Samples that produce hits to the majority of a genome’s coding regions may signify the abundance of a particular taxon. The complete genomes of highly prevalent species have been successfully constructed from complex metagenomic datasets, eg, the highly abundant crAssphage within fecal samples. Mapping reads to a reference genome of interest can also be used to classify and extract complete genome sequences. While neither approach has to date been applied to phage genome reconstruction within freshwater samples, some success has been possible with samples from other environments. The genomic plasticity of phages again presents a challenge as gene order and content between a reference genome and an isolate from nature may vary significantly, thus limiting the effectiveness of mapping strategies.

Furthermore, hypothesis-based inquiry of complex viral communities in nature is limited by the quality and quantity of data available. In an effort to gain insight into the putative predator–prey dynamics of phage and host, a metaviromic survey in parallel with a bacterial 16S rRNA gene survey of Lake Michigan nearshore waters was conducted. Metaviromic sequences were examined following the procedure outlined in Figure 1 (Velvet → getOrf → BLAST against NCBI’s viral RefSeq database). Metavirome contigs were classified via MetaVir; for each phage species detected, its annotated bacterial host was determined referencing the NCBI genome record and the associated literature. Figure 4 illustrates the observed (via 16S rRNA gene sequencing) relative abundances of bacterial phyla vs. the expected (via annotated host species for phage identified) for eight samples collected from one of the two sites along the Chicago shore of Lake Michigan. As Figure 4 shows, the majority of the phage species identified within the metaviromic analyses are documented for a host species that were not detected by the bacterial survey. Highly prevalent phage species included Pseudomonas-infecting species; however, pseudomonads were found in very low abundance based on the 16S rRNA gene survey. Most notably, Figure 4 suggests a large contingent of phages, which infect other bacterial taxa, predominately cyanobacteria; however, cyanobacteria were not detected via the complementary 16S rRNA gene survey.

The observed phage diversity reflects the underlying database (correlation of a number of predicted hosts to that of the annotated hosts for the available RefSeq phage sequences is $r^2 > 0.99$). While many factors (limited available data, misannotated and/or incomplete annotation of phage host species, etc.) may contribute to this discord, Figure 4 highlights the fact that without improved phage data resources, high-throughput sequencing projects are likely to uncover artifacts of the data repositories themselves. In many cases, the annotated host species for a phage genome is limited to a single, often laboratory, bacteria strain. Broad host range phages – phages capable of infecting bacteria belonging to the phyla Proteobacteria, Actinobacteria, and Bacteroidetes – have been isolated from Lake Michigan. This generalist lifestyle has benefits within environments for which bacterial communities, ie, susceptible host populations, are often ephemeral. It is important to note that the Lake Michigan broad host range viruses are relatives of the Pseudomonas PB1 phage, which to date has only been documented as infecting Pseudomonas spp. This suggests that phage species within the annotated collection of sequences may in fact have a broader host range than currently documented.
Bioinformatic approaches for metaviromic and bacteriophage analyses

Evolutionary Bioinformatics 2016:12(s1)

Moving Forward

In this review, we have illustrated the bioinformatic challenges faced by all metaviromic studies. These include poor representation of most virus groups in sequence repositories relative to their expected diversity in nature; overrepresentation of viruses infecting a handful of model laboratory bacterial strains; low rates of isolation and characterization of new phages in the laboratory; and generally, poor annotation of the genomes for phage that have been sequenced. Furthermore, while we have focused largely on dsDNA phages, RNA and ssDNA phages are important members of viral communities.73 While a few studies have been conducted,45,74,75 this fraction of the virome often goes unsequenced and unsampled due to additional difficulties with their genomic extraction and amplification. If the goal is to obtain a fair representation of all viruses in data repositories, then more work is also needed to identify the non-dsDNA phages. Even when sequenced, RNA phages can be especially difficult to identify, either due to their relative scarcity in a particular sample76 or due to biases in the analysis itself.77 Nevertheless, the bioinformatic obstacles highlighted here also extend to non-dsDNA phages. We have learned a great deal from metaviromic studies across biomes, and there are several steps available to improve the state of metaviromic bioinformatics going forward. We emphasize the role that freshwater studies can play in leading the way.

First, more studies of freshwater phages will help to improve the relative sampling of virus groups in sequence databases. To date, most studies have focused on phages from soil,78–80 sewage,81 and marine environments,10,36,82–84 with only a handful of studies reporting on the diversity of phages in freshwater (including those listed in Table 1). This is particularly surprising, as freshwater rivers and lakes have a direct impact on society and human health. Furthermore, freshwater ecosystems are available throughout the world at different latitudes and altitudes and include bodies of water of varying size, productivity, and geological history. Taken together with the seasonal dynamics of many freshwater systems, we expect these environments to harbor a great deal of undiscovered viral diversity.

Second, it is necessary to isolate and characterize phages uncovered from new metaviromic studies. Thus, virus groups in sequence repositories will be better represented aiding in the analyses of environmental viromes. Of central importance is the comprehensive characterization of new isolates. This goes beyond sequencing genomes; assaying growth characteristics (eg, latent period, generation time, and burst size) and the phage’s host range (following protocols such as Ref. 85) are imperative to furthering our understanding of environmental phages. In doing so, it will also be important to move beyond isolation on typical laboratory hosts and to include coisolated bacterial species. More work is also needed to characterize viral gene functions. Even for one of the best-studied laboratory strains, Enterobacteriophage T4, 114 of its 278 genes are currently annotated as hypothetical proteins in GenBank. Sequencing more genomes alone cannot fill in these gaps in our knowledge. RNA-seq has been used to track bacteriophage gene expression in vivo (eg, in the oral microbiome86), and similar approaches may also help to identify important, unannotated genes, in viral communities.

Figure 4. Phage–host populations within Lake Michigan nearshore waters.

Notes: Left: prevalence of bacterial phyla as determined by 16S rRNA gene sequencing.70 Right: expected prevalence of bacterial phyla based on the detection of phages annotated as infecting particular bacterial host species.66 The two sampling sites, along Chicago’s shoreline, are Montrose Beach (located 6.5 miles north of downtown Chicago) and 57th Street Beach (located 7.5 miles south of downtown Chicago). Sampling was conducted within the recreational beach waters.
Third, new analytical tools are necessary to make sense of the data that we do have. While the studies reviewed here often follow a bioinformatic analysis strategy akin to that presented in Figure 1, alternative approaches developed for prokaryotic and eukaryotic metagenomic studies (eg, those reviewed in the studies by Li and Homer and Nagarajan and Pop) may be applicable as well. Such methods require thorough vetting using both synthetic and real metavirome datasets. Furthermore, existing computational tools for protein structural and functional prediction may be adapted to the task of high-throughput prediction of novel functions in viral communities. New tools may also predict virus host range from genome sequences alone. Existing methods have recently been reviewed and it is found that simple methods, such as finding homologous genes in phage and bacteria using BLAST, can accurately predict many phage hosts. Though the best success rate of these current approaches is only near 40%, these tools provide an optimistic starting point for developing new methods. Improving host prediction from metagenomes would provide database-free analysis of local virus and bacterial interactions.

Freshwater systems, particularly freshwater lakes, offer unique opportunities for hypothesis-driven metaviromics that is to varying degrees independent of these database limitations. As discussed above, previous studies have explored how nutrient availability, seasonality, temperature, and human activity influence freshwater viral communities. Experimental metaviromics controlling for these variables should play an exciting role in both developing new computational tools and exploring viral ecology and evolution.

Acknowledgments
The authors thank the past and current members of the Putonti Lab who have assisted in collecting samples for the metaviromes generated by our group.

Author Contributions
Conceived and designed the study: JWS, SCW, CP. Collated and conducted the analysis of publicly available freshwater metavirome data sets: KB, KM, AC, ES. Wrote the first draft of the article: KB, KM, AC, ES. All the authors reviewed and approved the final article.

REFERENCES