Vol. 75: 117–128, 2015
doi: 10.3354/ame01752

AQUATIC MICROBIAL ECOLOGY
Aquat Microb Ecol

Published online June 4

Metagenomic assessment of viral diversity in Lake Matoaka, a temperate, eutrophic freshwater lake in southeastern Virginia, USA

Jasmin C. Green[1], Faraz Rahman[1], Matthew A. Saxton[2,3], Kurt E. Williamson[1,2,*]

[1]Department of Biology, College of William & Mary, Williamsburg, Virginia 23185, USA
[2]Environmental Science and Policy Program, College of William & Mary, Williamsburg, Virginia 23185, USA
[3]*Present address:* Department of Marine Sciences, University of Georgia, Athens, Georgia 30602, USA

ABSTRACT: Little is known about the composition and diversity of temperate freshwater viral communities. This study presents a metagenomic analysis of viral community composition, taxonomic and functional diversity of temperate, eutrophic Lake Matoaka in southeastern Virginia (USA). Three sampling sites were chosen to represent differences in anthropogenic impacts: the Crim Dell Creek mouth (impacted), the Pogonia Creek mouth (less impacted) and the main body of the lake (mixed). Sequences belonging to tailed bacteriophages were the most abundant at all 3 sites, with *Podoviridae* predominating. The main lake body harbored the highest virus genotype richness and included cyanophage and eukaryotic algal virus sequences not found at the other 2 sites, while the impacted Crim Dell Creek mouth showed the lowest richness. Cross-contig comparisons indicated that similar virus genotypes were found at all 3 sites, but at different rank-abundances. Hierarchical cluster analysis of multiple viral metagenomes indicated high genetic similarity between viral communities of related environments, with freshwater, marine, hypersaline, and eukaryote-associated environments forming into clear groups despite large geographic distances between sampling locations within each environment type. These results support the conclusion that freshwater viral communities are genetically distinct from virus assemblages in other environments.

KEY WORDS: Virus · Freshwater · Lake · Temperate · Metagenome · Sequencing · Virome

INTRODUCTION

Viruses are important and active components of the biosphere and can be found anywhere on the planet where cellular life exists. Through infection of microbial populations, viruses are significant drivers of ecological change in terms of host community composition and biogeochemical cycling due to cell lysis (Fuhrman 1999, Berdjeb et al. 2011), bacterial respiration rate due to lysis-induced changes in organic matter quality (Bonilla-Findji et al. 2008, Ram et al. 2013), and host metabolic capabilities due to horizontal gene transfer (Lindell et al. 2004, Kelly et al. 2013). The advent of metagenomics, which allows the sequencing of all the DNA present in an environmental sample without relying on cultures, has revolutionized the study of viral ecology and enabled unprecedented views of the taxonomic and functional diversity of the global virome.

While viral metagenomics has been used to investigate a wide variety of sample types including fermented foods (Park et al. 2011), hotsprings (Schoenfeld et al. 2008), and eukaryote-associated virus assemblies (Dinsdale et al. 2008, Cantalupo et al. 2011), the majority of studies have focused on marine samples (Breitbart et al. 2002, 2004, Angly et al. 2006). The current catalogue of viromes include relatively few freshwaters, and these tend to focus on

*Corresponding author: kewilliamson@wm.edu

exotic or extreme environments such as Antarctic lakes (López-Bueno et al. 2009), hypersaline lakes (Sime-Ngando et al. 2011), or ephemeral desert ponds (Fancello et al. 2013). Inland freshwaters provide a wide variety of ecosystem services, including (but not limited to) drinking water, irrigation, aquaculture, fish and shellfish production, recreation, wildlife habitat, and aesthetic value (Postel & Carpenter 2012). In spite of their importance, the viral communities of only 2 temperate freshwater lakes from a single study have been described using high-throughput DNA sequencing (Roux et al. 2012).

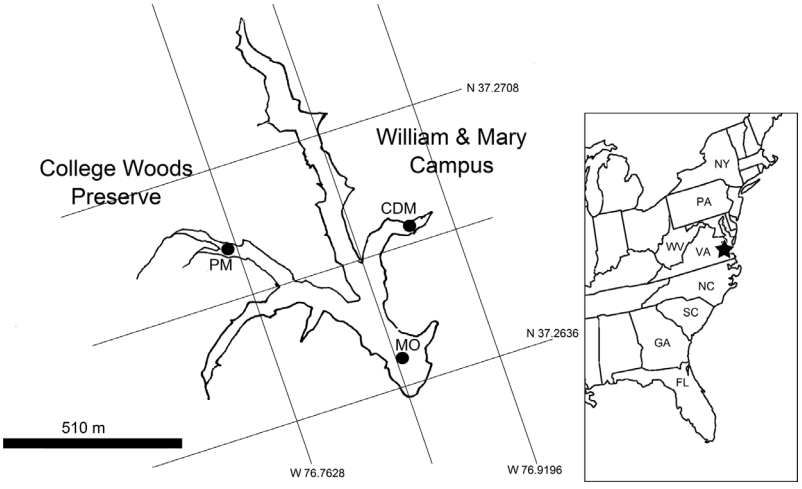

Fig. 1. Sampling sites at Lake Matoaka, Virginia, USA, and its watershed. MO, Matoaka open water; CDM, Crim Dell Mouth draining the William & Mary College campus (more impacted); PM, Pogonia Mouth draining the College Woods preserve (less impacted)

The goals of the present study were to investigate the viral community composition, taxonomic and functional diversity of Lake Matoaka, a temperate, eutrophic lake in southeastern Virginia, USA. In particular, 3 sampling sites were chosen to represent a gradient of inputs and anthropogenic impacts: a sub-watershed stream mouth draining unmanaged secondary growth forest, a sub-watershed stream mouth draining developed university campus, and the main body of the lake. Comparisons across these sites will provide insights as to how viral community composition may be influenced by human activities. Furthermore, comparing our results to other viral metagenomes will improve our understanding of how viral community structure varies by habitat type.

MATERIALS AND METHODS

Sampling sites

Lake Matoaka is a freshwater impoundment in Williamsburg, Virginia, USA, covering roughly 16 ha with an average depth of 2.5 m (max. depth 4.75 m) and water residence time of 75 d (Pensa & Chambers 2004). The lake was established by English colonists in 1718 through damming of the original creek system to form a mill-pond, making it the oldest man-made lake in Virginia and one of the oldest in the New World. Lake Matoaka is currently fed by 5 perennial streams and surface water runoff from its 600 ha watershed. College Creek to the north and Crim Dell Creek to the east are 2 major tributaries to the lake and are more heavily impacted by the nearby college campus and city of Williamsburg. Pogonia Creek and Strawberry Creek are 2 smaller streams that feed into the lake from the west (Fig. 1) and drain the minimally impacted ~150 ha College Woods preserve. The specific sample collection sites were the Crim Dell Mouth (CDM; 37.267° N, 76.721° W), the Pogonia Mouth (PM; 37.268° N, 76.727° W), and Matoaka open water (MO; 37.264° N, 76.722° W). The Crim Dell Creek sub-watershed has substantial development present (the College of William & Mary campus) while the Pogonia sub-watershed is dominated by secondary-growth forest. Crim Dell Creek and Pogonia Creek contribute water to Lake Matoaka at a base flow rate of 0.012 and 0.006 m^3 s^{-1}, respectively (http://media.wm.edu/content/as/kecklab/Weather/KeckWeather.htm).

Viral metagenome construction

Surface samples were collected from each site on 11 March 2013 in acid-washed polycarbonate bottles. A YSI-63 hand-held multimeter was used to measure temperature and conductivity and a YSI-55 hand-held probe was used to measure dissolved oxygen in the field. Samples were placed on ice and processed immediately upon return to the lab according to Thurber et al. (2009). Briefly, bacteria were removed by filtering 2 l samples through 0.22 µm and the filtrate was precipitated using polyethylene glycol (PEG-8000) (10% w/v) and NaCl (1 M) and incubated at 4°C overnight. Viral particles from the PEG pellet were purified using CsCl den-

sity gradient ultracentrifugation (1.7, 1.5, and 1.35 g ml^{-1} layers in SM buffer; virus particles collected from the 1.5–1.35 g ml^{-1} boundary) and DNAse treatment. Nucleic acids were extracted using the formamide procedure as previously described (Thurber et al. 2009). Viral nucleic acids were amplified using the Illustra Genomiphi V2 DNA amplification kit (GE Healthcare Life Sciences). Reactions were pooled and purified using the QIAgen DNeasy blood and tissue kit (QIAgen) before sequencing on a Roche Applied Sciences GS-FLX+ platform (454 Life Sciences).

Environmental metadata

Viral and bacterial abundance was determined using SYBR Gold-epifluorescence microscopy from triplicate subsamples (Noble & Fuhrman 1998, Hardbower et al. 2012). Virus-to-bacterium ratios were calculated based on averages of both viral and bacterial abundances for each sample. Bacterial production rate was determined based on ^{3}H-Leu incorporation rate from triplicate subsamples (Kirchman et al. 1985, Chin-Leo & Kirchman 1988). Nutrient levels (NO_2 + NO_3, NH_4 and total phosphorus) were determined using colorimetric assays with water filtered through Whatman glass fiber filters (GF/F) (Parsons et al. 1984). Chlorophyll *a* (chl *a*) was determined spectrophotometrically (Lorenzen 1967). Dissolved organic carbon was analyzed using GF/C (0.2 µm average pore size)-filtered water on a Shimadzu TOC-500. These data are summarized in Table S1 in the Supplement at www.int-res.com/articles/suppl/a075p117_supp.pdf.

Taxonomic and functional annotations

All libraries were dereplicated prior to analysis. The viral metagenome libraries were annotated using MG-RAST v. 3.3.6 (Meyer et al. 2008) and Metavir v. 2.0 (Roux et al. 2014) using an E-value cutoff of 10^{-5}, with MG-RAST accession numbers 4523574.3 (CDM), 4523575.3 (PM) and 4523576.3 (MO). MG-RAST generates taxonomic assignments based on BLASTx searches against the M5NR database (which includes SEED, KEGG, NCBI nr, Phantome, GO, EBI, JGI, UniProt, VBI, and eggNOG), and functional assignments based on BLASTx searches against the SEED-Subsystem database. Metavir generates taxonomic assignments of virus-related sequences based on BLASTx searches against RefSeq Virus database.

Contig assembly

Viral metagenomes were assembled using the GS De Novo Assembler v. 2.8 (Roche Diagnostics). Open reading frames (ORFs) were annotated in Metavir, which predicts ORFs for each contig through MetaGeneAnnotator and compares them to the RefSeq Virus protein database by BLASTp (E-value cutoff of 10^{-3}) and by HMMScan (Bit score cutoff of 30) to the PFAM database (Roux et al. 2014).

Viral community structure and diversity

Circonspect and GAAS (Genome relative Abundance and Average Size) (Angly et al. 2006, 2009) were installed on the SciClone computing cluster at the College of William & Mary (www.hpc.wm.edu/SciClone/Home) and run in the Unix environment. Results were exported to PHACCS (Angly et al. 2005) and MaxiPhi (Angly et al. 2006), which were run using MATLAB (R2014.a). Contig spectra were generated by Circonspect, which was run using 10 000 randomly selected 100 bp reads and generating assemblies with Minimo using the default parameters of 35 bp overlap with 98% similarity. Community structure and α-diversity were modelled in PHACCS using the contig spectra from Circonspect and average genome size from GAAS. Power law, exponential, logarithmic, and log-normal rank-abundance models were tested, and the best model was selected based on lowest error values. β-diversity was estimated with MaxiPhi by using cross-contig spectra from Circonspect, assuming 1000 permutations and a power-law community model. γ-diversity was estimated with PHACCS based on the mixed contig spectrum of the combined viral metagenomes (Angly et al. 2006).

RESULTS

Taxonomic distribution

A total of 48.4 Mbp were generated from the 3 sites (CDM, 11.6 Mbp; PM, 18.4 Mbp; MO 18.4 Mbp), corresponding to 123 481 individual reads (post-QC) with an average read length of 391 bp. Roughly 10% (10.6 to 10.9%) of reads failed quality checks based on Duplicate Read Inferred Sequencing Error Estimation (DRISEE; Keegan et al. 2012), k-mer profiles, and nucleotide bias within reads, and were removed from subsequent analysis (Fig. 2a). Annotation of

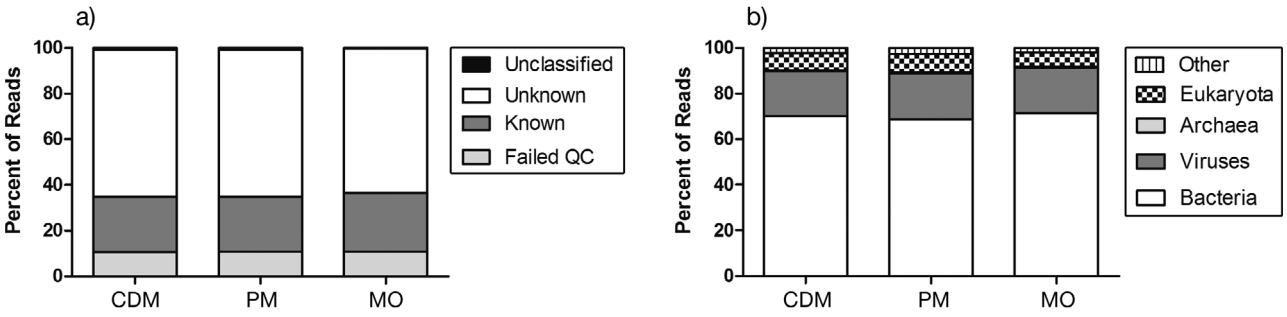

Fig. 2. (a) Classification of reads (Failed QC: did not meet minimum length/quality parameters; Known: predicted protein of known function; Unknown: predicted protein of unknown function; Unclassified: sequence contained no predicted open reading frame) and (b) domain of known reads according to best BLAST hit using MG-RAST (E-value cutoff of 10^{-5}) from 3 sites at Lake Matoaka and its watershed. Sites: CDM, Crim Dell Mouth; PM, Pogonia Mouth; MO, Matoaka open water

reads using MG-RAST indicated that 23.9 to 25.7% of reads had homology to known sequences, while over 60% had no known homologs in public databases (Fig. 2a). Of the reads with known homology, 19.5 to 19.9% were classified as viruses (Fig. 2b). Reads were also assessed using Metavir for in-depth analysis of library taxonomic composition using the GAAS tool (Angly et al. 2009). GAAS normalizes the number of hits according to genome size, generating a more accurate estimate of species abundances within a metagenome. According to the Metavir annotation, the majority of reads annotated as of viral origin belonged to dsDNA viruses and, among these, >79% matched with *Caudovirales*. Sequences belonging to *Podoviridae* were the most abundant in all metagenomes, followed by *Siphoviridae*, then *Myoviridae* (Table 1). All other identifiable viral families represented <1% of the total virome. The most represented viral genotype according to Metavir was *Puniceispirillum phage*, which ranged from 8.23 to 8.66% in each sample (Table 2). Only one eukaryotic virus, which infects the algae *Dunaliella viridis,* was

Table 1. Classification of reads hitting viral sequences: comparison of the taxonomic composition of water samples from Lake Matoaka, Virginia, USA, and its watershed. Analysis computed in Metavir via the GAAS (Genome relative Abundance and Average Size) tool (Angly et al. 2009) from a BLAST comparison with NCBI RefSeq complete viral genomes proteins using BLASTx (E-value < 10^{-5}). Sites: CDM, Crim Dell Mouth; PM, Pogonia Mouth; MO, Matoaka open water

Group	Order	Family	CDM (%)	PM (%)	MO (%)
dsDNA viruses,	–	*Adenoviridae*	0.14	0.34	0.03
no RNA stage	–	*Ampullaviridae*	0.00	0.02	0.00
	–	*Ascoviridae*	0.30	0.36	0.33
	–	*Baculoviridae*	0.02	0.01	0.00
Caudovirales	*Myoviridae*	9.36	10.11	8.78	
		Podoviridae	37.57	37.12	38.46
		Siphoviridae	28.51	28.19	28.73
		Unclassified *Caudovirales*	4.55	4.36	4.83
Herpesvirales	*Herpesviridae*	0.03	0.01	0.02	
		Alloherpesviridae	0.01	0.00	0.00
	–	*Iridoviridae*	0.06	0.02	0.04
	–	*Lipothrixviridae*	0.00	0.00	0.02
	–	*Mimiviridae*	0.06	0.09	0.06
	–	*Phycodnaviridae*	0.76	0.72	0.62
	–	*Poxviridae*	0.00	0.01	0.00
	–	*Rudiviridae*	0.02	0.02	0.02
	–	Unclassified dsDNA phages	13.62	13.90	12.69
	–	Unclassified dsDNA viruses	1.10	1.37	1.44
ssDNA viruses	–	*Inoviridae*	0.11	0.06	0.22
	–	*Microviridae*	0.00	0.08	0.00
	–	Unclassified ssDNA viruses	0.07	0.05	0.04
Unclassified phages	–	–	3.68	3.14	3.66
Unclassified virophages	–	–	0.04	0.03	0.00

detected at >1% abundance, and this virus was only found in the MO metagenome. Two additional virus genotypes were identified as unique to MO, *Sulfitobacter phage pCB2047-A* and *cyanophage KBS-P-1A* (Table 2).

Annotation of reads using MG-RAST indicated that 68.7 to 71.4% of known reads were classified as *Bacteria* (Fig. 2b). The most representative bacterial phylum in all samples was *Proteobacteria* followed by *Firmicutes* and *Bacteroidetes*, which comprised roughly 40%, 10%, and 7% of each library, respectively (Table S2 in the Supplement).

Table 2. The most represented viral genotypes for 3 sites at Lake Matoaka and its watershed. Viral genotypes computed via the GAAS (Genome relative Abundance and Average Size) tool (Angly et al. 2009) using BLASTx (E-value < 10^{-5}). Only viral genotypes comprising ≥1% of the total metagenome are shown. Hosts for all predicted viral species were bacteria, except for *Dunaliella viridis* virus SI2, which infects eukaryotic algae. Sites: CDM, Crim Dell Mouth; PM, Pogonia Mouth; MO, Matoaka open water

Metagenome Viral species	% total metagenome
CDM	
Puniceispirillum phage HMO-2011	8.66
Persicivirga phage P12024L	4.05
Pelagibacter phage HTVC010P	2.09
Celeribacter phage P12053L	1.94
Persicivirga phage P12024S	1.71
Marinomonas phage P12026	1.58
Flavobacterium phage 11b	1.57
Pseudoalteromonas phage RIO-1	1.54
Roseobacter phage SIO1	1.46
Cellulophaga phage phi38:1	1.37
Pelagibacter phage HTVC019P	1.31
Cellulophaga phage phi46:1	1.13
PM	
Puniceispirillum phage HMO-2011	8.24
Persicivirga phage P12024L	5.18
Cellulophaga phage phi38:1	1.84
Flavobacterium phage 11b	1.66
Persicivirga phage P12024S	1.66
Celeribacter phage P12053L	1.66
Pelagibacter phage HTVC010P	1.62
Roseobacter phage SIO1	1.30
Pelagibacter phage HTVC019P	1.21
Marinomonas phage P12026	1.21
Ralstonia phage RSK1	1.18
Pseudoalteromonas phage RIO-1	1.12
MO	
Puniceispirillum phage HMO-2011	8.23
Persicivirga phage P12024L	3.10
Pelagibacter phage HTVC010P	2.11
Celeribacter phage P12053L	1.82
Sulfitobacter phage pCB2047-A	1.62
Pelagibacter phage HTVC019P	1.55
Ralstonia phage RSK1	1.51
Roseobacter phage SIO1	1.45
Pseudoalteromonas phage RIO-1	1.30
Persicivirga phage P12024S	1.27
Marinomonas phage P12026	1.19
Cellulophaga phage phi38:1	1.13
Cyanophage KBS-P-1A	1.13
Dunaliella viridis virus SI2	1.05

Functional annotation

Metabolic subsystems were annotated using MG-RAST, based on the best BLASTx hit against the SEED-subsystems database. About 6% (6.2 to 6.7%) of reads could be classified by function in this way. The most represented subsystem in each sample was related to phages, prophages, and transposable elements, accounting for 40.5 to 43.7% of each metagenome. Other highly represented (>5%) functional categories included nucleoside and nucleotide metabolism; clustering-based subsystems; cofactors, vitamins, prosthetic groups, and pigments; and DNA metabolism (Fig. 3). Pairwise comparisons of the metabolic profiles of each library using *t*-tests showed that subsystems related to respiration were significantly more highly represented ($p < 0.01$) in PM as compared to MO. The representation of carbohydrate metabolism was significantly higher ($p < 0.05$) in PM as compared to CDM (Fig. 3).

Contig assembly and analysis

A total of 1986 contigs were assembled (454 CDM, 801 PM, and 731 MO), representing 25.6 to 29.3% of reads. The average contig size across all libraries was 1.28 kbp, with the largest contigs being 29.8 to 34.8 kbp long (Table S3 in the Supplement). Contigs were annotated using Metavir, which compares sequences to the RefSeq Virus database using BLASTp (E-value threshold of 10^{-3}). Contigs were adenine-thymidine (AT)-rich (54.7 to 58.5%) with a high proportion of predicted open reading frames with no match in any database (ORFans), roughly 85% of ORFs across all contigs (Table S4 in the Supplement). The ORFs with known function coded for recognizable phage structural, genome packaging, and nucleotide metabolism genes. Contigs were also mapped against the most abundant reference genome using the contig analysis tools available through Metavir. The reference genome with the highest coverage in each of the 3 libraries was *Puniceispirillum phage HMO-2011*, with genome coverage of 51.5 to 55.9% (Table S5 in the Supplement).

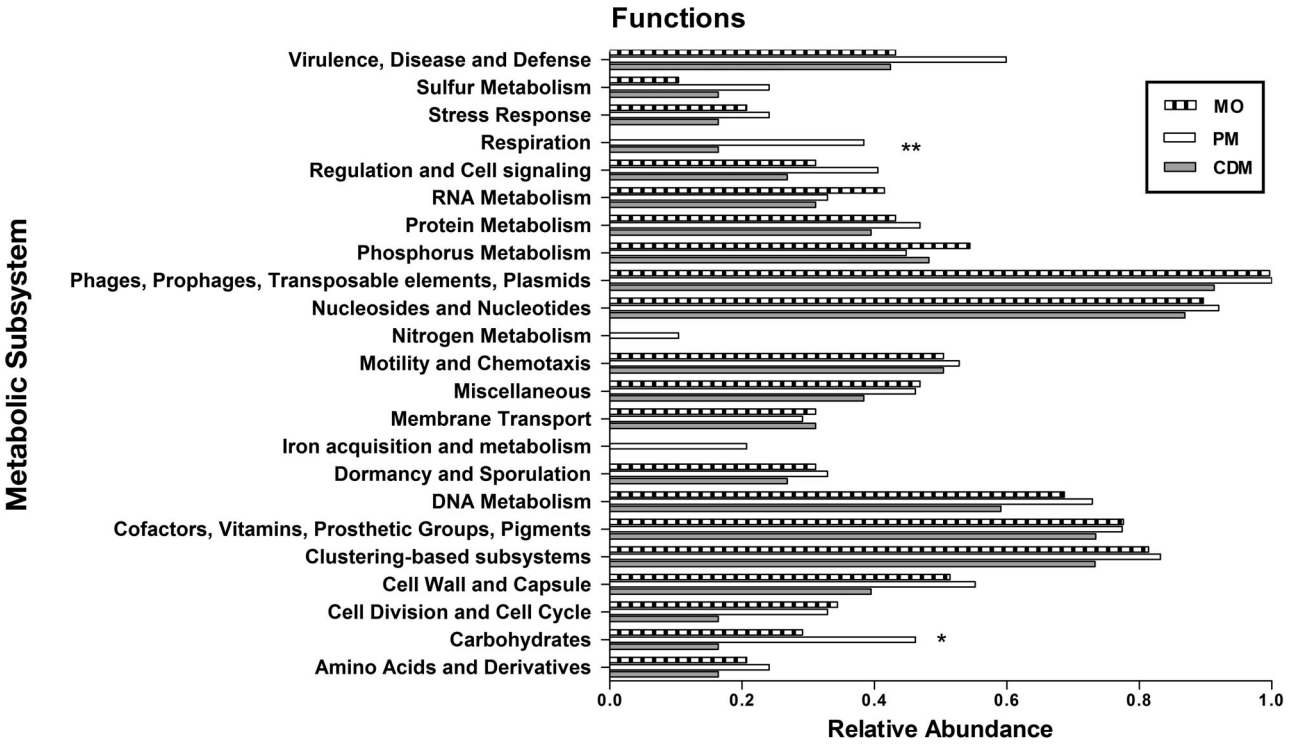

Fig. 3. Relative abundance of reads from samples taken at Lake Matoaka and its watershed assigned to functional subsystems using MG-RAST. Classification based on best BLAST hit (E-value cutoff of 10^{-5}) using M5NR database. Asterisks indicate metabolic functions with significantly different ($p < 0.05$) representation across the lake metagenomes based on pairwise comparisons (*t*-tests). Sites: CDM, Crim Dell Mouth; PM, Pogonia Mouth; MO, Matoaka open water

Community structure and diversity

Based on PHACCS results, the community structures of the Lake Matoaka viral metagenomes — defined by richness (R), evenness (E), and Shannon-Wiener diversity index (H') — were graphically represented as rank-abundance curves and best modeled by the power law (Fig. S1 in the Supplement). MO had the highest richness and overall diversity (R = 7662, H' = 8.66) followed by PM (R = 6622, H' = 8.46), and then CDM (R = 4956, H' = 8.17) (Table 3). The γ-diversity represented by all 3 lake viral communities was estimated at 7360 genotypes with H' of 8.58 (Table 3). To measure species similarity between the viral communities, the β-diversity was estimated using cross-contig spectra (Angly et al. 2005). This approach compares the sequences in common and the changes in the rank-abundance of genotypes (percent shared and percent permuted, respectively) between viromes without relying on database matches. According to the MaxiPhi results, the 3 sites were highly similar in terms of shared viral genotypes but with large differences in rank-abundances of genotypes across sites (Fig. 4). At 92% similarity, MO and CDM had the lowest overlap of genotypes between any pair of sites, but the low degree of permutation suggests that shared genotypes were found at similar rank-abundances. By contrast, MO and PM were highly similar with 99.7% of viral genotypes shared across the 2 sites, but the larger degree of permutation (3.87%) suggests that the same genotypes exist at different rank abundances within the 2 samples. This trend is especially pronounced in comparing the 2 stream mouths: while both sites appear to share similar viral genotypes, the high degree of permutation (9.18%) indicates that genotypes that

Table 3. Comparison of viral community structure and diversity at Lake Matoaka, and its watershed according to PHACCS analysis. Parameters: average shotgun sequence length 850 bp, overlap 20 bp, max. 100 000 genotypes; model, Power Law. H', Shannon-Wiener diversity index. Sites: CDM, Crim Dell Mouth; PM, Pogonia Mouth; MO, Matoaka open water; Mixed, aggregate of all 3 sites

	CDM	PM	MO	Mixed
Avg. genome size (bp)	59 538	63 132	62 914	61 861
Error	25.49	83.02	10.27	13.54
Richness	4956	6717	7662	7360
Evenness	0.961	0.886	0.968	0.964
Most abundant genotype (%)	0.915	0.716	0.594	0.711
H'	8.17	8.46	8.66	8.58

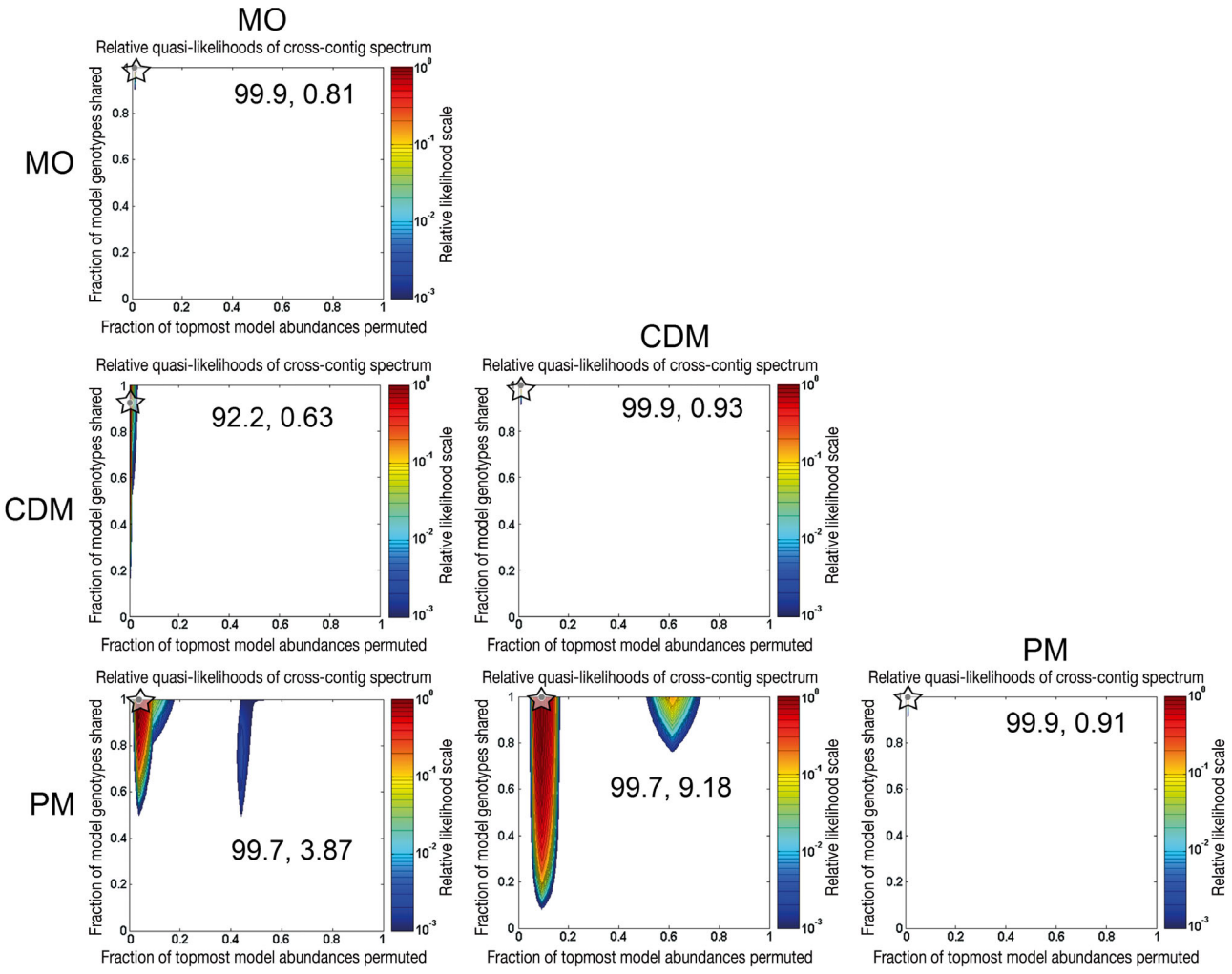

Fig. 4. Pairwise comparisons and estimation of beta diversity (as determined by MaxiPhi) of the metagenomes sampled at Lake Matoaka and its watershed. For each plot, *y*-axis values show the fraction of modeled genotypes shared between any pair of viromes; *x*-axis values show the fraction of topmost model abundances permuted; colors indicate likelihood scale; stars indicate maximum likelihood. Numerical values within each plot show estimated percentage of viral genotypes shared (first value) and estimated percent permuted (second value). Sites: MO, Matoaka open water; CDM, Crim Dell Mouth; PM, Pogonia Mouth

may be highly abundant in one stream are relatively rare in the other (Fig. 4).

Comparison with other freshwater viral metagenomes

The 3 Lake Matoaka viral metagenomes were compared to the viromes of other freshwater studies, including 2 French lakes (Roux et al. 2012), 4 Saharan gueltas (Fancello et al. 2013), 3 aquaculture ponds (Rodriguez-Brito et al. 2010), and 4 waste water treatment plant samples (Tamaki et al. 2012). Three marine viral metagenomes (Angly et al. 2006) and a Chesapeake Bay virome (Bench et al. 2007) were also included in the comparison. Sequences were classified according to best BLASTx hit against the M5NR database and multiple comparisons were visualized using principal coordinate analysis (PCoA) on the MG-RAST server. Metagenomes clustered broadly according to habitat, with marine samples clustering in the bottom right quadrant and the large lakes and waste water treatment plant samples gathering in top left quadrant (Fig. S2 in the Supplement). The viromes from the present study clustered closely together and were grouped in the top-right quadrant with the aquaculture and Chesapeake Bay samples. PM was separated slightly from MO and CDM, in the direction of the large lake and waste water treatment plant samples.

To obtain a broader comparison independent of reference databases, our freshwater viral metagenomes

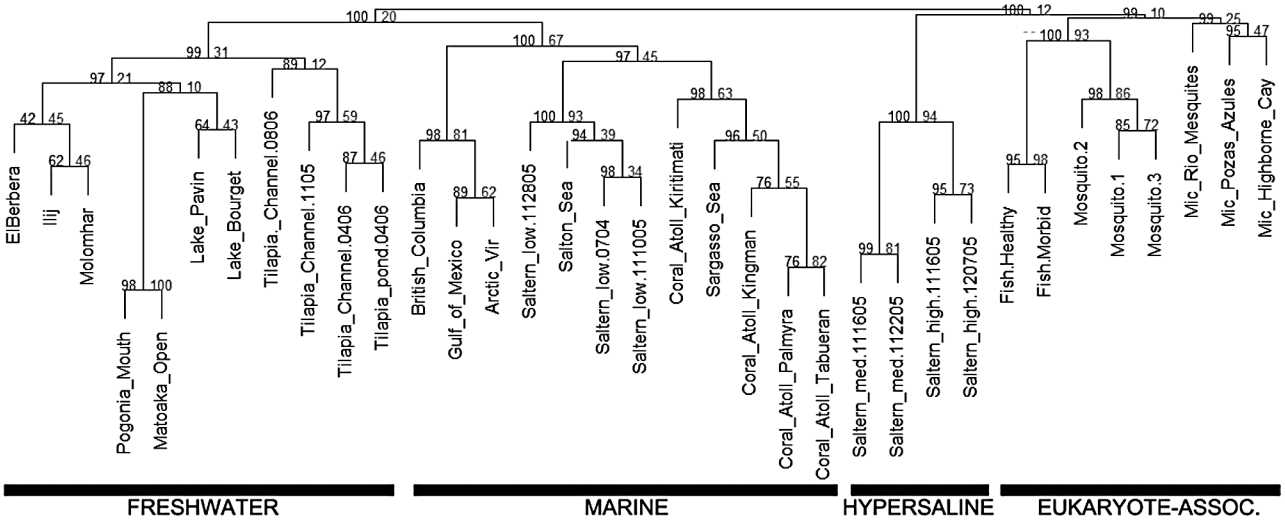

Fig. 5. Hierarchical cluster tree based on tBLASTx comparisons across viral metagenomes from this study and other studies representing a wide variety of water body types, geographical locations, and environments. Numerical values to the right of each branch point indicate approximately unbiased p-values; values to the left indicate bootstrap probability. Samples shown are from this study (temperate lake, Virginia, USA), Saharan Gueltas (El Berbera, Ilij, Molomhar; Fancello et al. 2013), temperate lakes in France (Lake Bourget, Lake Pavin; Roux et al. 2012), tilapia aquaculture (California, USA; Rodriguez-Brito et al. 2010), marine water column (Bay of British Columbia, Gulf of Mexico, Arctic Sea, Sargasso Sea; Angly et al. 2006), solar salterns (California, USA; Rodriguez-Brito et al. 2010), corals (Dinsdale et al. 2008), microbialites (Mexico; Desnues et al. 2008), tilapia-associated (Rodriguez-Brito et al. 2010) and mosquito-associated (Ng et al. 2011)

were compared to each other, as well as 33 other viral metagenomes encompassing multiple sample types (Dinsdale et al. 2008, Rodriguez-Brito et al. 2010) using tBLASTx. The resulting hierarchical clustering tree, drawn in R using the pvclust package (distance method = correlation, cluster method = average, bootstrap number = 10 000), indicated a grouping of viral communities according to 4 main sampling environments: eukaryote-associated, hypersaline waters, marine and low-salinity waters, and freshwaters (Fig. 5). Within the freshwater cluster, 4 sub-groups were observed, consisting of: (1) samples from Saharan gueltas, (2) samples from Lake Matoaka, (3) samples from larger French lakes (Pavin and Bourget), and (4) samples from aquaculture ponds.

DISCUSSION

The development of tools to sequence and analyze metagenomes has revolutionized the study of viral ecology. Since the advent of metagenomics, viruses have been studied in widely varying systems from the open ocean to fermented foods (Angly et al. 2009, Park et al. 2011). However, the viral communities of relatively common inland freshwaters such as small, temperate lakes are underrepresented in the literature. Currently, viral metagenomes have been analyzed for 2 freshwater lakes: Lake Bourget (mesotrophic, 4450 ha) and Lake Pavin (oligotrophic, 44 ha), both located in France (Roux et al. 2012). No studies have yet explored viral composition of eutrophic lakes, which encompass approximately half of all inland freshwater lakes in the US (US EPA 2009). In the present study, 3 viral metagenomes from Lake Matoaka, representing the mouths of 2 feeding streams and the main body of the lake, were compared to provide a fine-grained view of the viral communities present in a eutrophic freshwater system. To our knowledge, it is the first metagenomic study of a small (16 ha), temperate, eutrophic lake.

Methodological considerations and limitations

As with most viral metagenome studies, a 0.22 µm filtration step was included during sample processing to remove cells. Large viruses were likely removed during this process, resulting in the low representation of *Mimiviridae* (≤0.01%) and *Phycodnaviridae* (≤0.76%) in the present study, despite these families being common in other aquatic environments (Ghedin & Claverie 2005). Another concern is the use of GenomiPhi to amplify the viral DNA to ensure enough starting product for sequencing. GenomiPhi uses a phi29-derived DNA polymerase,

which has been shown to preferentially amplify small ssDNA (Angly et al. 2006, López-Bueno et al. 2009, Roux et al. 2012). Despite this, our libraries exhibited a high representation of dsDNA viruses (>96%, Table 1), which suggests that any quantification bias caused only minor variations in the results. Roughly 70% of known reads were classified as bacteria, which is consistent with previous viral metagenomes (Angly et al. 2006, Rodriguez-Brito et al. 2010, Roux et al. 2012, Fancello et al. 2013). Inspection of viral concentrates using epifluorescence microscopy revealed no evidence of bacterial contamination. The high number of bacterial hits is likely due to a number of factors, including database bias towards bacterial sequences and horizontal gene transfer between viruses and microbial hosts, allowing viruses to gain bacterial host genes and vice-versa (Seguritan et al. 2003, Sullivan et al. 2005).

Comparison of Lake Matoaka viral metagenomes

Water samples from the mouths of 2 tributaries and the open water of Lake Matoaka were selected for this study. Of the sequences returned, over 60% had no known homologs in public databases (Fig. 2a), supporting the idea that much of the global viral diversity remains unknown. Generally, the taxonomies of the 3 Lake Matoaka metagenomes were similar, being dominated by dsDNA tailed bacteriophages (Tables 1 & 2). Upon deeper inspection of phage genotypes, viruses known to infect common freshwater bacteria, particularly *Flavobacteria* and *Proteobacteria* (Zwart et al. 2002), were the most represented in all 3 samples (Table 2). The fact that cyanophages and viruses infecting eukaryotic algae were found at abundances >1% only in MO represented a noticeable difference between the 3 libraries (Table 2). However, it is reasonable that the hosts for these viruses would be found at greater abundances in the open water, where they are able to receive direct sunlight for photosynthesis as opposed to the more shaded stream mouths. Indeed, this interpretation is supported by the higher chl *a* concentration at MO as compared to the other sites (Table S1).

Analysis of the β-diversity of these samples also revealed that each sample contained similar viral genotypes but with large differences in the representation of each genotype across sampling sites. This is consistent with the results of a previous study (Rodriguez-Brito et al. 2010) that suggest that samples from the same ecosystem (in this case, a freshwater lake) tend to display a characteristic taxonomy. The lower degree of overlap between CDM and MO (92%, Fig. 4) is consistent with the greater disparity in genotype richness between the 2 sites. There may have been viral genotypes present in the richer MO community that simply were not found in the less-diverse CDM community (Table 3). However, the relatively large volume of water flow from CDM may explain the similar rank-abundances across these 2 sites since CDM is a larger feeder of MO and contributes more to the overall microbial community composition of the lake. By contrast, while PM exhibits different rank-abundance of viral genotypes from MO, the relatively small proportion of water flow that PM provides into the lake means that PM contributes much less to the microbial community composition of the lake as a whole (Site MO). The comparison between the PM and CDM communities is most interesting. The high degree of overlap suggests that the same fundamental set of viral genotypes are found in both sub-watersheds. However, the large degree of permutation (9.18%, Fig. 4) suggests that differences in land use patterns have significant impacts on the representation of particular genotypes within the community.

In terms of metabolic capabilities, it is interesting that the PM library showed significantly higher representation of genes associated with respiration and carbohydrate metabolism as compared to MO and CDM, respectively (Fig. 3). While the PM sample had the lowest bacterial and viral abundances of the 3 sites, the bacterial production rate was almost twice as high as that measured in the MO sample, and 10-fold higher than that measured in the CDM sample (Table S1). These differences in bacterial production rate likely stem from the quantity and quality of carbon compounds available within each site. Viruses may acquire different metabolic genes which provide selective advantage during infection under particular growth requirements imposed upon the host community (Lindell et al. 2004, Kelly et al. 2013). Thus, it is possible that viruses at PM have acquired a higher proportion of genes associated with respiration and carbohydrate metabolism based on selective pressures placed upon the host community at this site.

According to PHACCS analysis, MO displayed the highest richness and overall diversity while CDM had the least (Table 3). One explanation for the observed differences in viral diversity is that each of the stream mouth sites receive inputs from their respective streams while MO receives many inputs from creeks and surface runoff from the College Woods basin. In this regard, the main body of the lake represents a sort of mixing chamber that includes

viruses that have entered the lake from various sources, as well as viruses that have been produced *in situ*. This interpretation is supported by the fact that the viral genotypes from MO exhibited a large degree of overlap with both stream mouths (Fig. 4). At the same time, the larger representation of cyanophages and algal viruses at MO (Table 2) reflects the likely production of these viruses at that particular site. The Crim Dell Mouth, which exhibited the lowest viral richness, is also subject to the greatest human impact. It receives regular nutrient inputs via runoff from the campus grounds, as well as additional inputs from road and parking lot runoff, and refuse from the campus community. In contrast, the Pogonia stream is surrounded by ~480 ha of undeveloped oak-hickory forest, and is subject to much lower human impact. This finding agrees with other studies indicating that ecosystems that are highly impacted by human activity tend to display lower biodiversity compared to less impacted areas (Rodrigues et al. 2013, Goldenberg Vilar et al. 2014). Recent studies have indicated that this trend extends to microbes including bacteria and fungi (Kebli et al. 2012, Rodrigues et al. 2013), and one previous report suggested that viral richness is likewise impacted (Fancello et al. 2013). The results of the present study support the hypothesis that an inverse relationship between degree of human impact and species richness also applies to viruses.

Comparison with other aquatic communities

Estimated richness is highly variable across different viral metagenomes. With an average richness of ~6000 viral genotypes, Lake Matoaka falls in the middle of the spectrum among currently characterized freshwater systems. The expansive, mesotrophic Lake Bourget shows an order of magnitude greater richness (~40 000 genotypes) while the average richness of desert gueltas (perennial freshwater ponds) was about an order of magnitude lower (~400 genotypes) (Roux et al. 2012, Fancello et al. 2013). The proportion of unknown sequences in Lake Matoaka is typical of other viromes, with fewer than 30% sequences returning database matches. The samples from Lake Matoaka were generally AT-rich with a high proportion of ORFans (Table S4). These features appear to be characteristic of viral metagenomes and increase our confidence that the sequences are viral in origin, despite the high percentage of unknowns.

The Lake Matoaka viromes were all dominated by tailed, dsDNA bacteriophages from the order *Caudovirales*. While the desert gueltas were also dominated by *Caudovirales*, albeit to a lesser extent (Fancello et al. 2013), this trend is not common to all freshwater viromes. In the 2 large freshwater lakes in France, the ssDNA microvirus and circovirus groups were the main constituents, encompassing 73 to 92% of the libraries (Roux et al. 2012). It is possible that the representation of circoviruses within a library depends partially upon the version of the GenomiPhi kit used for a particular library preparation. However, these specific details are not available for the Lake Bourget and Lake Pavin libraries for comparison. No doubt multiple technical differences in virome preparations can bias the eventual representation of ssDNA viruses. Additionally, a study that analyzed the seasonal differences in viral communities in an Antarctic lake found that ssDNA viruses were more dominant in the spring whereas dsDNA viruses were more abundant in the summer (López-Bueno et al. 2009). Since most freshwater viral metagenome projects, including the present study, have relied upon single time-point samples, it remains an open question whether or not other lakes exhibit such temporal shifts in viral community composition.

Comparison of aquatic viral metagenomes based on PCoA (Fig. S2) indicated a general grouping according to sample type. However, this analysis is limited to sequences with matches in reference databases. Based on hierarchical clustering in which complete viromes were compared to each other via BLASTx, viral communities appear to group according to environment type, with viral metagenomes from freshwater samples forming a distinct cluster (Fig. 5). This reflects significant genetic similarity between these viral assemblages, in spite of large geographic distances between sample locations (eastern North America, western North America, Europe, and Africa). These results also reinforce similar findings from an independent research group (Roux et al. 2012).

CONCLUSIONS

We found that the viral communities of a temperate, eutrophic freshwater lake were of intermediate richness (~5000 to 7500 genotypes) when compared to other freshwater viromes. The sub-watersheds of the lake and open lake sites shared over 90% of the same viral genotypes, but differed greatly in the rank abundance of genotypes by site. This suggests that the lake watershed shares a common core of virus genotypes but that land use practices can impact the rank-abundance of genotypes within a given site.

Our results supported the hypothesis that anthropogenic impacts can reduce species richness and that this trend extends to viral communities. Finally, hierarchical clustering of multiple viromes supported the idea that freshwater viral communities are genetically distinct from other virus assemblages.

Acknowledgements. We thank John Busch and Joseph Jones at Selah Genomics, Greenville, SC, for assistance with library construction and pyrosequencing. We also thank Tom Crockett and Eric J. Walter in the College of William & Mary Information Technology/High Performance Computing Group for their time and assistance with installing and troubleshooting Ciconspect, MaxiPhi, and PHACCS. Thanks also to Dr. Florent Angly for his assistance with successful operation of the bioinformatics analysis packages above. Thanks to Dr. Randolph Chambers and Tim Russell at the Keck Environmental Field Laboratory, College of William & Mary, for water chemical analyses. This work was performed in part using computational facilities at the College of William & Mary which were provided with the assistance of the National Science Foundation, the Virginia Port Authority, and Virginia's Commonwealth Technology Research Fund. This work was funded by an Andrew W. Mellon Postdoctoral Fellowship in Environmental Science to M.A.S. and a grant from the Jeffress Memorial Trust (J-988) to K.E.W. The authors declare no conflict of interest.

LITERATURE CITED

Angly F, Rodriguez-Brito B, Bangor D, McNairnie P and others (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41–50

Angly FE, Felts B, Breitbart M, Salamon P and others (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368

Angly FE, Willner D, Prieto-Davó A, Edwards RA and others (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLOS Comput Biol 5:e1000593

Bench SR, Hanson TE, Williamson KE, Ghosh D, Radosovich M, Wang K, Wommack KE (2007) Metagenomic characterization of Chesapeake Bay virioplankton. Appl Environ Microbiol 73:7629–7641

Berdjeb L, Pollet T, Domaizon I, Jacquet S (2011) Effect of grazers and viruses on bacterial community structure and production in two contrasting trophic lakes. BMC Microbiol 11:88

Bonilla-Findji O, Malits A, Lefevre D, Rochelle-Newall E, Lemee R, Weinbauer MG, Gattuso JP (2008) Viral effects on bacterial respiration, production and growth efficiency: consistent trends in the Southern Ocean and the Mediterranean Sea. Deep-Sea Res II 55:790–800

Breitbart M, Salamon P, Andresen B, Mahaffy JM and others (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci USA 99:14250–14255

Breitbart M, Felts B, Kelley S, Mahaffy JM, Nulton J, Salamon P, Rohwer F (2004) Diversity and population structure of a near-shore marine-sediment viral community. Proc R Soc Lond Ser B Biol Sci 271:565–574

Cantalupo PG, Calgua B, Zhao G, Hundesa A and others (2011) Raw sewage harbors diverse viral populations. MBio 2:e00180

Chin-Leo G, Kirchman DL (1988) Estimating bacterial production in marine waters from the simultaneous incorporation of thymidine and leucine. Appl Environ Microbiol 54:1934–1939

Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S and others (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452: 340–U5

Dinsdale EA, Edwards RA, Hall D, Angly F and others (2008) Functional metagenomic profiling of nine biomes. Nature 452:629–632

Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D, Desnues C (2013) Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. ISME J 7:359–369

Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399:541–548

Ghedin E, Claverie JM (2005) Mimivirus relatives in the Sargasso Sea. Virol J 2:62

Goldenberg Vilar A, van Dam H, van Loon EE, Vonk JA, van Der Geest HG, Admiraal W (2014) Eutrophication decreases distance decay of similarity in diatom communities. Freshw Biol 59:1522–1531

Hardbower DM, Dolman JL, Glasner DR, Kendra JA, Williamson KE (2012) Optimization of viral profiling approaches reveals strong links between viral and bacterial communities in a eutrophic freshwater lake. Aquat Microb Ecol 67:59–76

Kebli H, Brais S, Kernaghan G, Drouin P (2012) Impact of harvesting intensity on wood-inhabiting fungi in boreal aspen forests of Eastern Canada. For Ecol Manag 279: 45–54

Keegan KP, Trimble WL, Wilkening J, Wilke A and others (2012) A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. PLoS Comput Biol 8(6): e1002541

Kelly L, Ding H, Huang KH, Osburne MS, Chisholm SW (2013) Genetic diversity in cultured and wild marine cyanomyoviruses reveals phosphorus stress as a strong selective agent. ISME J 7:1827–1841

Kirchman D, K'Nees E, Hodson R (1985) Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. Appl Environ Microbiol 49:599–607

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci USA 101:11013–11018

López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A (2009) High diversity of the viral community from an Antarctic lake. Science 326:858–861

Lorenzen CJ (1967) Determination of chlorophyll and phaeopigments: spectrophotometric equations. Limnol Oceanogr 12:343–346

Meyer F, Paarmann D, D'Souza M, Olson R and others (2008) The metagenomics RAST server — a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386

Ng TFF, Willner DL, Lim YW, Schmieder R and others (2011) Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoSONE 6: e20579

Noble RT, Fuhrman JA (1998) Use of SYBR Green I for rapid epifluorescence counts of marine viruses and bacteria. Aquat Microb Ecol 14:113–118

Park EJ, Kim KH, Abell GCJ, Kim MS, Roh SW, Bae JW (2011) Metagenomic analysis of the viral communities in fermented foods. Appl Environ Microbiol 77:1284–1291

Parsons TR, Maita Y, Lalli CM (1984) A manual of chemical and biological methods for seawater analysis. Pergamon Press, Oxford

Pensa MA, Chambers RM (2004) Trophic transition in a lake on the Virginia coastal plain. J Environ Qual 33:576–580

Postel S, Carpenter S (2012) Freshwater ecosystem services. In: Daily G (ed) Nature's services: societal dependence on natural ecosystems. Island Press, Washington, DC, p 195–214

Ram ASP, Palesse S, Colombet J, Sabart M, Perriere F, Sime-Ngando T (2013) Variable viral and grazer control of prokaryotic growth efficiency in temperate freshwater lakes (French Massif Central). Microb Ecol 66:906–916

Rodrigues JLM, Pellizari VH, Mueller R, Baek K and others (2013) Conversion of the Amazon rainforest to agriculture results in biotic homogenization of soil bacterial communities. Proc Natl Acad Sci USA 110:988–993

Rodriguez-Brito B, Li L, Wegley L, Furlan M and others (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4:739–751

Roux S, Enault F, Robin A, Ravet V and others (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS ONE 7:e33641

Roux S, Tournayre J, Mahul A, Debroas D, Enault F (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis. BMC Bioinformatics 15:76

Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, Mead D (2008) Assembly of viral metagenomes from Yellowstone hot springs. Appl Environ Microbiol 74:4164–4174

Seguritan V, Feng IW, Rohwer F, Swift M, Segall AM (2003) Genome sequences of two closely related *Vibrio parahaemolyticus* phages, VP16T and VP16C. J Bacteriol 185:6434–6447

Sime-Ngando T, Lucas S, Robin A, Tucker KP and others (2011) Diversity of virus-host systems in hypersaline Lake Retba, Senegal. Environ Microbiol 13:1956–1972

Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. PLoS Biol 3:e144

Tamaki H, Zhang R, Angly FE, Nakamura S and others (2012) Metagenomic analysis of DNA viruses in a wastewater treatment plant in tropical climate. Environ Microbiol 14:441–452

Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F (2009) Laboratory procedures to generate viral metagenomes. Nat Protoc 4:470–483

US EPA (2009) National lakes assessment: a collaborative survey of the nation's lakes. US Environmental Protection Agency, Washington, DC, EPA 841–R–09–001

Zwart G, Crump BC, Agterveld MPK, Hagen F, Han S (2002) Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. Aquat Microb Ecol 28:141–155

Editorial responsibility: Gunnar Bratbak, Bergen, Norway

Submitted: December 15, 2014; Accepted: March 20, 2015
Proofs received from author(s): May 22, 2015